Understanding the Epidemiology of Bluetongue virus in South India using statistical models

Mohammed Mudassar Chanda

Linacre College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2015

Understanding the Epidemiology of Bluetongue virus in South India using statistical models

Mohammed Mudassar Chanda

Linacre College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2015

Abstract

Bluetongue is an economically important midge-borne disease affecting domestic and wild ruminants worldwide. The disease, caused by the bluetongue virus (BTV), is highly endemic in South India, occurring with varying severity every year since 1963, causing high morbidity and mortality, resulting in huge economic losses to subsistence farmers, impacting the GDP of the country and affecting food security. The bluetongue epidemiological system in South India is characterized by an unusually wide diversity of susceptible ruminant hosts, many potential *Culicoides* vector species and numerous pathogen serotypes and strains. These factors (intrinsic and extrinsic) contribute to disease impacts that vary widely over geographical space

Chapter 2, deals with identification of remote sensed variables in discriminating between presence and absence of bluetongue outbreaks and development of risk map using non-linear discriminant analysis (NLDA) approach.

Chapter 3 deals with understanding the role of extrinsic factors such as monsoon conditions in driving seasonality in BTV outbreaks over two decades in Andhra Pradesh, India using a Bayesian Poisson regression model framework, accounting for temporal autocorrelation.

In chapter 4 the mean annual numbers of outbreaks in each district in South Indian states were examined in relation to land-cover, host availability and climate predictors using a Bayesian generalised linear mixed model with Poisson errors and a conditional autoregressive error structure to account for spatial autocorrelation.

In Chapter 5 the annual number of outbreaks in each district in South India was examined in relation to climate predictors (temperature and precipitation) at different lags using a Bayesian generalized linear mixed model with Poisson errors.

In chapter 6 a range of suitable predictors was considered for identifying their relationships with bluetongue outbreaks using Bayesian Network Modelling (BNM), and the important variables were used to develop a Bayesian geostatistical model accounting for spatial autocorrelation

The analysis resulted in development of spatial risk maps at district and village level, district level yearly predictions and monthly state level predictions, which can contribute to the development of an early warning system for the disease in South India.

Acknowledgements

I would like to express deep sense of gratitude to my Supervisors Prof David J Rogers and Dr. Beth Purse for their constant support, motivation and encouragement. It has been a learning experience under their able guidance and supervision. I thank them for their patience and support during difficult situations. I would have not be able to complete my thesis without their suggestions, constructive criticism and moral support.

I would like to thank past and present members of the Spatial Ecology and Epidemiology Group, including Dr. Sarah Randolph, Dr.Luigi Sedda, Dr.David Morley, Dr. David Benz, Dr. Nick Golding and Dr. Elinor Godfrey for their valuable suggestions and help during my entire stay in the department. Besides the SEEG group, I would like to thank Dr. Kate Searle for her help during my stay at CEH, Edinburgh

Thanks are due to Dr. Gaya Prasad (ADG, ICAR) for his help in getting official leave to pursue my PhD and also for his valuable suggestions and guidance. A special thanks to my former Director of NIVEDI, Dr.Prabhudas for his initial help in applying for this fellowship and present Director Dr. H. Rahman for supporting me throughout my PhD and providing the data for bluetongue outbreaks. I also thank my senior colleague Dr. M.R.Gajendragad for his support and guidance. Special thanks to my senior colleague at ICAR-NIVEDI, Dr. D.Hemadri for supporting me during initial stages of my career and encouraging to take up this fellowship.

I am thankful to Principal Investigator of the project, Dr. Simon Carpenter for providing this opportunity to work in a multidisciplinary project. I also thank other project partners; Dr. Minaxi Prasad, Dr. S.M. Byregowda, Dr. YNR Reddy, Dr. YKM Reddy for their suggestions and their help in collating data on bluetongue outbreaks. I thank my funders BBSRC, DFID, CIDLID, and Scottish Government for the PhD studentship and ICAR

I would like express my deep sense of gratitude to all my teachers for teaching me at different stages of my career. I thank all my friends who have been constant support especially Dr. Jagadish Hiremath for his motivation and suggestions.

Lastly, I want to thank all my family members for encouraging me to take up this project especially my elder brother Khizer Chanda, younger brother Mufassir Chanda and my sister Dr.Afshan Shireen for their moral support. I thank my brother in law Dr. Azhar Akhil, for motivating me to apply for this fellowship. I also thank Tanveer’s brothers Nazim and Taj for supporting us during difficult times.

I want to thank my mother for her support, encouragement and prayers. Finally, I would like to dedicate my thesis to my wife, Tanveer Jahan for her love, sacrifices, encouragement and support throughout my PhD. It would have been impossible for me to complete my PhD without her support.

Contents

List of tables vi

List of figures...x

1. General Introduction 1

1.1 Bluetongue virus, clinical signs and transmission 2

1.2 History of bluetongue 4

1.3 Bluetongue in India 4

1.3.1 Clinical BT in South India versus seroprevalence in North India 4

1.3.2 Impacts of BT on rural livelihoods, food security and economic losses 6

1.4 Intrinsic factors 8

1.4.1 Role of past introduction of crossbreeds to improve local breeds 8

1.4.2 Herd immunity 9

1.4.3 Serotypes 10

1.4.4 Breed and host susceptibility 10

1.4.5 Bluetongue virus serotypes variability and re-assortment.11

1.4.6 Stress of other diseases, lambing and other stressors 12

1.5 Extrinsic factors 12

1.6 How can statistical models help in understanding intrinsic and extrinsic drivers of disease at different spatial and temporal scales? ..13

1.6.1 Presence and absence methods 13

1.6.2 Time series analysis 15

1.6.3 Spatial analysis of count data..18

1.6.4 Space-time analysis of epidemiological data 23

1.6.5 Bayesian methods 25

1.7 Aims and structure of the thesis 26

2. A Non-Linear Discriminant Analysis approach to distinguish areas of presence and absence of Bluetongue in South India 28

2.1 Introduction..29

2.2 Materials and methods..33

2.2.1 Bluetongue presence and absence data.............................33

2.2.2 Remotely sensed variables...35

2.2.3 Non-Linear Discriminant Analysis Model (NLDA) description..36

2.3 Results..44

2.3.1 Single best model results..45

2.3.2 Averaged bootstrap model results.....................................47

2.4 Discussion..57

3. Role of intrinsic and extrinsic factors in driving temporal patterns in bluetongue outbreaks in India: a Bayesian time series regression approach 62

3.1 Introduction..63

3.1.1 Extrinsic factors...64

3.1.2 Intrinsic factors...65

3.1.3 Forecasting and Early warning system.............................66

3.2 Materials and methods...67

3.2.1 Disease data..67

3.2.2 Weather and El-Niño data..68

3.2.3 Statistical methods..69

3.3 Results..77

3.4 Discussion...93

4. Understanding spatial variation in occurrence of bluetongue outbreaks between districts in South India using a Bayesian Poisson regression model 98

4.1 Introduction..99

4.1.1 Diversity of climate, hosts, and landcover on BTV outbreaks and vector distribution..100

4.1.2 Spatial autocorrelation..102

4.2 Materials and Methods...104

4.2.1 Bluetongue outbreak data 104

4.2.2 Land-cover data 105

4.2.3 Host species and sheep breed data 106

4.2.4 Rainfall and temperature data 106

4.2.5 Modelling approach 108

4.2.6 Model building and selection of predictors 109

4.3 Results...112

4.3.1 South India model 112

4.3.2. State-level models 114

4.4 Discussion and Conclusion 127

5. Understanding the inter-annual spatio-temporal risk factors for bluetongue outbreaks across South India using Bayesian Poisson regression modelling..133

5.1 Introduction...134

5.1.1 Space-time analysis of epidemiological data 137

5.2 Materials and Methods..141

5.2.1 Disease data 141

5.2.2 Predictor variables 141

5.2.3 Modelling approach 144

5.3 Results...152

5.3.1 Spatio-temporal patterns in bluetongue outbreaks...........152

5.3.2 Comparison of space-time models.................................152

5.4 Discussion 170

6. Village level spatial analysis of bluetongue cases using Bayesian Network modelling and Bayesian model based geostatistics 175

6.1 Introduction...176

6.1.1Variable selection methods 177

6.1.2 Bayesian Network Modelling 178

6.1.3 Bayesian geostatistical model 182

6.2 Materials and methods 186

6.2.1 Bluetongue case data for Andhra Pradesh 186

6.2.2 Climatic predictors 186

6.2.3 Land cover and topographical predictors 187

6.2.4 Livestock census data 187

6.2.5 Bayesian network modelling 188

6.2.6 Bayesian geostatistical modelling approach 193

6.3 Results 197

6.3.1 Bayesian Network Modelling results 197

6.3.2 Restricted DAG model results (Fig 6.6) 202

6.3.3 Spatial regression (Bayesian geostatistical) model results 203

6.4 Discussion 219

7. Discussion and Conclusions 225

7.1 Introduction 226

7.2 Different epidemiological systems in South India 227

7.3 Intrinsic and extrinsic factors in determining bluetongue variability 228

7.3.1 Intrinsic and extrinsic factors determining temporal variability of BTV outbreaks at state and district level 228

7.3.2 Role of Intrinsic and extrinsic factors in spatial variability of BTV outbreaks 231

7.3.3 Space-time methods to understand intrinsic and extrinsic factors 232

7.4 Variable selection 235

7.5 Predictions at different spatial and temporal scales to help in controlling BTV in South India 236

7.5.1 Global early warning systems for livestock diseases 236

7.5.2 NADRES model for forecasting bluetongue in India 237

7.5.3 Temporal predictions 238

7.5.4 District level spatial analysis of bluetongue outbreaks....239
7.5.5 District level annual predictions of BTV outbreaks 239
7.5.6 Lead times for forecasts..240
7.5.7 Stakeholders using risk maps and predictive models 241
7.5.8 Risk communication to stakeholders...............................242
7.6 Summary and conclusions 243

References...245

List of tables

Table 2.1: Accuracy and validation accuracy statistics (Kappa, sensitivity and specificity) for the three model series.. 50

Table 2.2: Mean values of the top ten ranked variables from the best bluetongue model with lowest AIC_C among the 100 bootstrap models. 51

Table 2.3: Key to variable names for the best model among the 100 bootstrap models. 52

Table 2.4: Mean ranks of the top ten variables from all the 100 bootstrap models in the three Model series, 1, 2 and 3. 53

Table 2.5: Model accuracy matrix of the best model (pseudo absence with 3 presence and absence clusters each) among 100 bootstrap models. 55

Table 2.6: Overall accuracy statistic for the best model among 100 bootstrap models. The overall kappa accuracy of this model is 0.86 and AUC: 0.9987. 55

Table 3.1: Selection of monthly models (rainfall and temperature) and combination of rainfall and temperature variables based on reduced DIC. 89

Table 3.2: Mean co-efficient and credible interval of the best model (rainfall at lag 2, plus AR (1)) 90

Table 3.3: Root mean square error (RMSE) and DIC for rainfall at lag 2 with AR (1), rainfall at lag 2 and AR (1) only model. 90

Table 4.1: Land cover variables considered in the combined South India model and individual state level models. 116

Table 4.2: Host and breed variables considered in the combined South India model and individual state level models. 117

Table 4.3: Deviance information criterion (DIC) for the best models of bluetongue severity across South India, where predictors are drawn from a single category of environmental predictors 121

Table 4.4: Final model of bluetongue severity across south India which includes the best combination of host, climate and landscape predictors (top row). 122

Table 4.5: Mean coefficient values and credible intervals for fixed effects environmental predictors which describe the average annual number of bluetongue outbreaks per district for the best model for South India. 123

Table 4.6: Mean coefficient values and credible intervals for fixed effects environmental predictors which describe average annual number of bluetongue outbreaks in Andhra Pradesh for the best subset models across the land-cover, climate and host categories. 125

Table 4.7: Mean coefficient values and credible intervals for fixed effects environmental predictors which describe average annual number of bluetongue outbreaks in Karnataka for the best subset models across the land-cover, climate and host categories. 126

Table 4.8: Mean coefficient values and credible intervals for fixed effects environmental predictors which describe average annual number of bluetongue outbreaks in Tamil Nadu for the best subset models within the land-cover, climate and host categories. 127

Table 5.1: Deviance information criterion (DIC) and correlation co-efficient for the best models, where predictors are drawn from a single category of environmental predictors.. 159

Table 5.2: Mean coefficient values along with their standard deviation(sd) and credible intervals for fixed effects of environmental predictors with spatial and temporal effects (structured and unstructured components)which describe sum of annual number of bluetongue outbreaks per district. 160

Table 5.3: best model with and without interaction along with their DIC’s, pD (effective number of parameters), and the squared correlation- co-efficient (r2) between observed and predicted data, and Root mean square error on the test data 161

Table 5.4: Mean coefficient values along with their standard deviation(sd) and credible intervals for fixed effects of environmental predictors with spatial and temporal effects (structured and unstructured components) which describe the sum of annual number of bluetongue outbreaks per district (values in bold indicate that they are significant because the CIs do not bridge zero) using a negative binomial distribution model. 162

Table 5.5: Negative binomial model with and without interaction along with their DIC's, pD (effective number of parameters), and the squared correlation- co-efficient (r2) between observed and predicted data. 163

Table 6.1: Variables used in the Bayesian network analysis. The variable labels used in the DAGs. 205

Table 6.2: Log marginal likelihood (network score) of DAG's with different parent limits and with banned relationships. 206

Table 6.3: Credible intervals of variables associated with the bluetongue node. 210

Table 6.4: Credible intervals of variables associated with host nodes (in bold) identified in the best DAG with a six parent limit (unrestricted). 210

Table 6.5: Credible intervals of variables associated with land cover nodes (in bold) identified in the best DAG with a six parent limit (unrestricted model). 211

Table 6.6: Credible intervals of variables associated with the terrain or climate nodes (in bold) identified in the best DAG with a six parent limit. 212

Table 6.7: DIC of different models fitted using two clustering methods to generate 4 different knots in the observations. In medoid clustering the knot must be located at one of the actual observations. 217

Table 6.8: Mean coefficients and their credible intervals of the Bayesian spatial regression model (Equation 1) with 256 knots using the k-medoids based algorithm. 217

Table 6.9: Root mean square error (RMSE) statistics for the training and test data for the Bayesian spatial regression model with 256-knots-k-medoids............. 218

List of figures

Figure 1.1 Clinical signs observed in sheep affected with bluetongue 2

Figure 1.2 Map of India highlighting the three South Indian states under study (Andhra Pradesh, Karnataka and Tamil Nadu) and the other states. 6

Figure 1.3: Time line of the bluetongue outbreaks in the past, and crossbreeding of local breeds with exotic sheep in India. 9

Figure 2.1: Maps of potential explanatory discriminating variables affecting presence and absence of bluetongue in South India. 49

Figure 2.2: Graphical representation ('rainbow plots') of the ranks of predictor variables to show how often any particular variable was selected across the 100 bootstrap models for the Model 1, 2 and 3 series of models. 54

Figure 2.3: Risk map for bluetongue in south India using Fourier processed MODIS variables 56

Figure 2.4: (a) Distribution of clusters in one of the Model 3 series of models, with 3 presence and 3 absence clusters. 57

Figure 3.1: Time series plots of (a) monthly Bluetongue outbreaks (b) Monthly rainfall (mm) (c) Maximum temperature (^{0}C) and (d) sea surface temperature data(^{0}C) from 1992-2009 (El-Niño 3). 81

Figure 3.2: Box plots of the raw monthly and yearly mean bluetongue outbreaks and environmental variables. (A) Mean monthly and yearly bluetongue outbreaks (B) Mean monthly and yearly Rainfall (C) Mean monthly and yearly Maximum temperature. 82

Figure 3.3: Plots of (A) Autocorrelation function (B) Partial autocorrelation function of bluetongue outbreaks. 83

Figure 3.4: Plots of (A) Autocorrelation function (B) Partial autocorrelation function of maximum temperature and rainfall. 84

Figure 3.5: Cross-Correlation functions of bluetongue outbreaks with rainfall, maximum, mean and minimum temperatures. (A–D) Cross-correlation functions (CCF) with (A) Maximum temperature (B) Mean temperature, (C) Minimum temperature and (D) Rainfall.. 85

Figure 3.6: Dominant frequencies in the monthly bluetongue outbreaks time series & monthly rainfall time series and their cross-correlation. 86

Figure 3.7: Dominant intra-annual frequencies in the monthly bluetongue outbreaks time series & monthly El Niño 3 time series and their cross-correlation. 87

Figure 3.8: correlation between El Niño 3 with maximum temperature (A) and rainfall (B) Cross-wavelet coherence (correlation) 88

Figure 3.9: Plot of best model fit with rainfall at lag 2 and the AR(1) error term. Grey areas correspond to the 95% credible interval, red solid line is the predicted number of outbreaks and the circles are the observed number outbreaks. 91

Figure 3.10: Plot of model fits with A) rainfall at lag 2 only (no AR(1)) and B) AR(1) only (no rainfall at lag 2). 92

Figure 3.11: Plot of observed bluetongue outbreaks and predictions on “out of fit” data with best model with lowest DIC (Rainfall at lag 2 with AR (1)).............. 93

Figure 4.1: Maps of climate variables in southern India: (A) Annual mean temperature (^{0}C) (B) Annual average North-East monsoon precipitation (mm) and (C) Average South West monsoon precipitation (mm) 118

Figure. 4.2 Maps of district level land use variables in southern India: (A) Rain fed croplands; (B) closed (>40%) broadleaved deciduous forest (>5m) and (C) Water bodies 119

Figure. 4.3 Maps of district level host abundance in southern India (all on a log10 scale): (A) non-descript sheep (log.$_e$); (B) buffalo (log.$_e$); and (C) crossbred cattle (log.$_e$) (obtained from National livestock census 2007). 120

Figure. 4.4 (A) Observed average annual number of outbreaks in South India from PD_ADMAS; and (B) Predicted average annual number of outbreaks from the best model. 124

Figure 5.1: Variable selection procedure: Best variables were selected from individual category (land cover, climate and host) and the final model was selected by fitting different combinations of selected predictors.................................. 151

Figure 5.2: Bluetongue outbreaks in South India (1992-2009).The outbreak data were log.$_e$ transformed. 157

Figure 5.3: Bluetongue outbreaks from 1995 to 2006 in different districts of south India (1-79).The X-axis indicates different years and Y-axis indicates log$_e$ (natural logarithm) transformed BTV outbreaks. 158

26Figure 5.4: Comparison of the observed (A) average BTV outbreaks (1992-2007) with (B) fitted outbreaks using best model with spatial and temporal random effects. 164

Figure 5.5: observed and predicted bluetongue outbreaks using the model from best combination of variables with spatial and temporal random effects on the “out of fit” data for two years (2008 & 2009)... 165

Figure 5.6: Wavelet power spectra of- the time series of annual rainfall in three Indian states & of the annual sea surface temperature in the eastern Pacific region. 166

Figure 5.7: Correlation (Cross-wavelet coherence) between dominant frequencies in the annual rainfall time series & annual sea surface temperature (El-Nino 3) time series.. 167

Figure 5.8: Correlation (Cross-wavelet coherence) between dominant frequencies in the South-West monsoon rainfall time series & annual sea surface temperature (El-Nino 3) time series.. 168

Figure 5.9: Correlation (Cross-wavelet coherence) between dominant intra-annual frequencies in the North-East monsoon rainfall time series & annual sea surface temperature (El-Nino 3) time series. 169

Figure 6.1: Semi-variogram showing different parameters; Nugget, range and sill 183

Figure 6.2: Depiction of a Bayesian Graphical model (DAG). 189

Figure 6.3: Flow chart explaining steps for Bayesian Network Modelling and Bayesian geostatistical method. 197

Figure 6.4: The final DAG with a six parent limit (unrestricted). 207

Figure 6.5: The final DAG with six parent limit (unrestricted) after bootstrapping (>50% retained in the bootstrapping) 208

Figure 6.6: The DAG (restricted or banned model) with six parent limit. 209

Figure 6.7: Distribution of different independent variables in Andhra Pradesh (AP) used in the analysis (A) logged (natural logarithm) Indigenous cattle (B)logged (natural logarithm) non-descript sheep population (C) area covered by post-flooding or irrigated croplands (D) area covered by rain fed croplands (E) Annual mean temperature (^{0}C) and (F) South-West monsoon rainfall (mm). 213

Figure 6.8: (A) Observed maximum bluetongue cases (natural logarithm) and (B) Total sheep population (natural logarithm) in Andhra Pradesh. 214

Figure 6.9: Semivariogram on the residuals of the non-spatial model. Dots indicates the empirical semi-variogram. 215

Figure 6.10: Plot of the locations of centroids used to cluster the villages of AP for measurement of semi-variance and spatial structure. 216

Figure 6.11: Observed (A) and fitted (B) bluetongue outbreaks using Bayesian spatial regression model (Equation 1) with 256 knots using the k-medoids based algorithm. 218

Figure 7.1: Influence of different intrinsic and extrinsic factors on bluetongue outbreaks in South India across different spatial and temporal scales. 234

Chapter 1

General Introduction

1.1 Bluetongue virus, clinical signs and transmission

Bluetongue virus (BTV) is a double-stranded RNA virus (family Reoviridae, genus Orbivirus) that causes bluetongue in ruminants. Clinical signs include fever, nasal discharge, excessive salivation, facial odema, ulceration and cyanosis of tongue (bluetongue), coronitis, skeletal muscle damage (Fig. 1.1). Severe clinical signs are observed in certain breeds of sheep, especially European fine wool and mutton breeds and mild symptoms in cattle, goat, camelids and carnivores (Maclachlan et al., 2009; Taylor, 1986; Verwoerd & Erasmus, 2004). Currently, twenty six BTV serotypes are recognized globally (Maan et al., 2012).

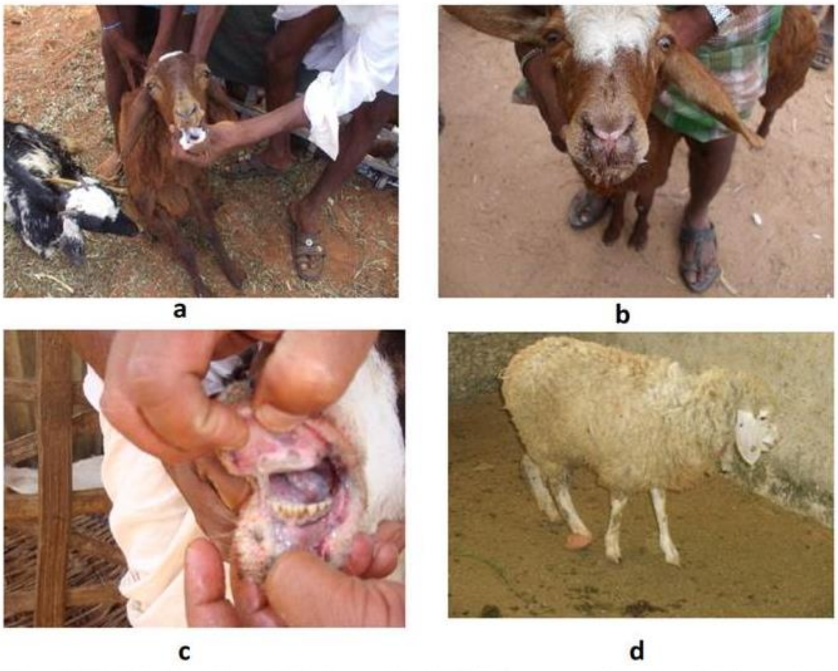


Figure 1.1 Clinical signs observed in sheep affected with bluetongue: Bluetongue is characterized by (a) high fever, salivation (b) oedema of face, lips and mucopurulent discharge (c) ulceration, haemorrhages of oral mucous membranes, cyanosis of tongue and (d) severe lameness in the later part of the disease due to coronitis. (Pictures taken during 2009 BTV outbreaks in Karnataka, India. Courtesy: ICAR-NIVEDI, Bangalore, Karnataka, India).

Bluetongue virus is transmitted between susceptible hosts by certain species of biting midges, *Culicoides* (Diptera : Ceratopogonidae) (Mellor, 2000) and is hence restricted to areas and seasons where competent adult vector populations occur, broadly in the tropical and sub-tropical countries between 35^{0}S and 40^{0}N (Tabachnick, 2003) but extending in places (Europe, N. America) up to 50^{0}N (Purse et al., 2005; Saegerman et al., 2008). *Culicoides* populations build up to high abundances under suitable conditions, and adults can disperse over a few to tens of kilometres a day, leading to rapid disease spread (Burgin et al., 2013; Sedda et al., 2012). The rapidity of the spread of bluetongue and its high impact on trade led to its designation as an OIE-listed, notifiable disease (Gibbs & Greiner, 1994).

There are 1,357 known species of *Culicoides* of which only 20 are of medical or veterinary importance (Purse et al., 2015). Almost all species (96%) of *Culicoides* are obligate blood suckers of mammals and birds (Mellor et al., 2000). The life cycle of *Culicoides* consists of the egg, four larval stages, pupal and adult stages (Mellor et al., 2000). Sixty one species of *Culicoides* have been reported from India (Sen & Gupta, 1959; Wirth & Hubert, 1989). The actual number may be greater, but there is a lack of systematic vector studies in the sub-continent (Ilango, 2006).

Culicoides imicola is the principal vector transmitting BTV in tropical and sub-tropical countries (Mellor et al., 2000). It is also the prinicpal vector in some parts of Southern Europe. Palaearctic species (at least one species each from the *Obsoletus* and *Pulicaris* groups) can also act as effective vectors of BTV (Mellor et al., 1990; Mellor & Prrzous, 1979; Wilson &

Mellor, 2009). *C.sonorensis* is the prinicpal vector in Northern America (Tabachnick, 2003; Wittmann et al., 2002), whereas *C.brevitarsis* is the main vector in Australia (Muller et al., 1982).

1.2 History of bluetongue

Based on clinical signs, bluetongue was first recognized in the year 1902 and was known as Malarial Catarrhal fever or Epizootic Catarrhal fever affecting imported European breeds of sheep in Africa (Henning, 1949; Maclachlan et al., 2009). Initially (1902-1943) bluetongue was restricted to Africa. The first outbreak outside Africa was reported in Cyprus in 1943 (Gambles, 1949). The disease was thereafter reported from the Middle East (1949), the Americas (1952), Europe (1956), the Indian subcontinent (1959) and Australia (1975) (Erasmus et al., 2009; Verwoerd & Erasmus, 2004). Major outbreaks occured in Europe (Spain and Portugal) from 1956-1960 and have re-occurred over a wider area since 1998 (Wilson & Mellor, 2009).

1.3 Bluetongue in India

1.3.1 Clinical BT in South India versus seroprevalence in North India

Since the first confirmed outbreak of bluetongue in India in 1963 (Sapre, 1964) outbreaks of BTV in sheep have been reported from many different states including Gujarat, Haryana, Jammu & Kashmir, Karnataka, Andhra Pradesh and Tamil Nadu. The disease is highly endemic, with clinical outbreaks reported regularly in the Southern states (Fig. 1.2) (Prasad et al., 2009). Elsewhere in India there are many reports (Bhalodiya & Jhala, 2002; Bhanuprakash et al., 2007; Chauhan et al., 2004; Prasad et al., 1992) of sero-prevalence in different species of livestock (Dadawala et al., 2013; Desai, 2004; Jain et al., 1992), but on an irregular basis. In a large scale national sero-survey of antibodies against bluetongue virus involving twelve states (around 7000 serum samples screened), approximately 50% of the samples from sheep tested positive and 58% from goats. The past reports of widespread high sero-prevalence from different parts of India but with clinical disease restricted to South India suggests either there is a diversity of environmental factors (climate, host and land cover) which might be playing a role in driving this heterogeneity in bluetongue outbreaks or there is underreporting in the rest of the Indian states (Ahuja et al., 2008).

Although virus isolations have been made from different species of *Culicoides* (Dadawala et al., 2012), there are no studies in India of the vectorial capacity or seasonality of any of the *Culicoides* species (*C.brevitarsis, C.actoni, C.peregrinus, C.imicola, C.oxystoma, C.fulvus, C.brevipalpis*) (Patel et al., 2007; Reddy & Hafeez, 2008), some of which are proven candidate vectors in other countries (Mellor et al., 2000; Muller et al., 1982).

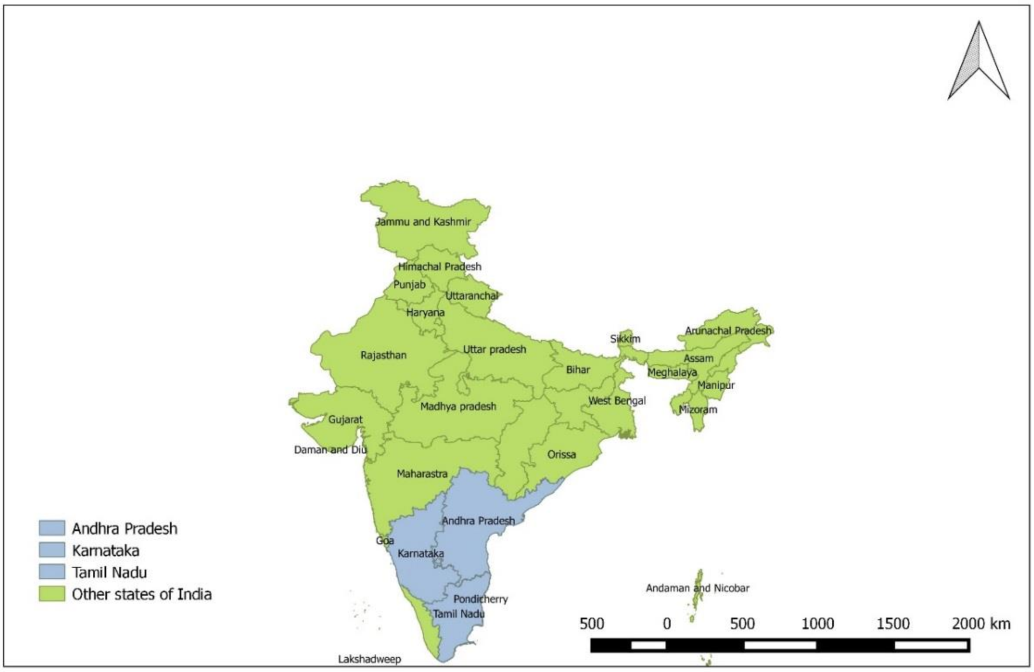


Figure 1.2 Map of India highlighting the three South Indian states under study (Andhra Pradesh, Karnataka and Tamil Nadu) and the other states.

1.3.2 Impacts of BT on rural livelihoods, food security and economic losses

Bluetongue in India was ranked as the top disease of sheep between 1997 and 2005, with 2313 outbreaks resulting in approximately 0.4 million cases and 64,086 deaths (Ahuja et al., 2008). Bluetongue losses are both direct (up to 30% case fatality rates) (Sreenivasulu et al., 2003) and indirect, from animal weight loss, poor wool quality and trade restrictions.

South Asian countries (including India) account for 22% of the world's human population and more than 40% of the world's poor (World Bank, 2001, IFAD, 2001). Eighty percent of the poor live in rural areas. India accounts for over 75% of the South Asian population, with 82% of poor people, and contributes 70% (excluding poultry) of the livestock population of the region. Sheep farming contributes 7.5% (65 million sheep) and goat farming 11.6 % (126 million goats) of the global sheep and goat populations respectively (FAOSTAT, 2008). The livestock sector contributes around 6% of India's Gross Domestic Product (Ali, 2007) and 25% of agricultural GDP and has grown at an annual rate of 5.6% over the last few decades (Ali, 2007). This sector supports the livelihoods of over 200 million rural poor (Ahuja et al., 2008) with around five million households engaged in rearing small ruminants. The majority of these are small, marginal and landless households that are able to rear small ruminants due to the low initial investment and operational costs involved (Mcleod & Kristjanson, 1999). Mixed farming systems, with complementary crop and livestock management activities, thus cover 83% of all agricultural land in India (Chacko et al., 2010).

In addition to its food and manure production functions, livestock rearing increases household resilience in capital terms, for times of crisis (Ahuja et al., 2000; World Bank, 1999). Several empirical studies indicate that livestock rearing has a positive impact on equity in terms of income, employment and poverty reduction in rural areas (see references cited in Ali, 2007). Meat and milk consumption are both set to grow by 2020 (Delgado et al., 1999), representing a significant opportunity for India to boost rural incomes and accelerate the pace of poverty reduction, but disease represents a significant barrier to realising the productive potential of livestock (Ahuja et al., 2008).

Bluetongue causes huge mortality and affects the productivity of the animals in India and other affected countries, but disease impacts vary widely in space and time even in endemic areas The two main groups of drivers that may control disease dynamics are intrinsic and extrinsic (also known as endogenous and exogenous) (Koelle & Pascual, 2004; Ruiz et al., 2006).

1.4 Intrinsic factors

1.4.1 Role of past introduction of crossbreeds to improve local breeds

In endemic regions, local ruminant breeds often demonstrate resistance to disease, and clinical signs are often observed only in imported breeds that are more susceptible to infection (Coetzee et al., 2012; Daniels et al., 2003), or when the virus is introduced into new regions with immunologically naive and susceptible sheep population. There is no documented evidence of BTV or bluetongue-like disease in India before the confirmed report in 1963 (Sapre, 1964), therefore an attempt was made to search the animal disease outbreaks database (http://digital.nls.uk/indiapapers/browse/) from 1900 onwards. The search resulted in one report of unusual sheep mortality in the year 1935 and, interestingly, a bacterial or parasitic cause was ruled out. This rare episode of high sheep mortality was preceded, in 1930, by the first documented evidence of crossbreeding to improve the wool quality of local breeds in India.

During the two decades following the 1963 confirmed report of BTV in India, incidents of recorded transmission were confined to imported breeds of sheep, particularly Southdown, Rambouillet, Russian Merino, Corriedale and Suffolk. Crossbreeding in India, using imported Rambouillet sheep, was started in 1962 with the establishment of the CSWRI (Central Sheep and Wool Research Institute). Subsequently the AICRP (All India Co-ordinated Research Project) was founded in 1970, with different centres all over India. From 1981, BTV was also observed in local and crossbred sheep (Prasad et al., 2009). The disease now occurs every year in South India with varying severity.

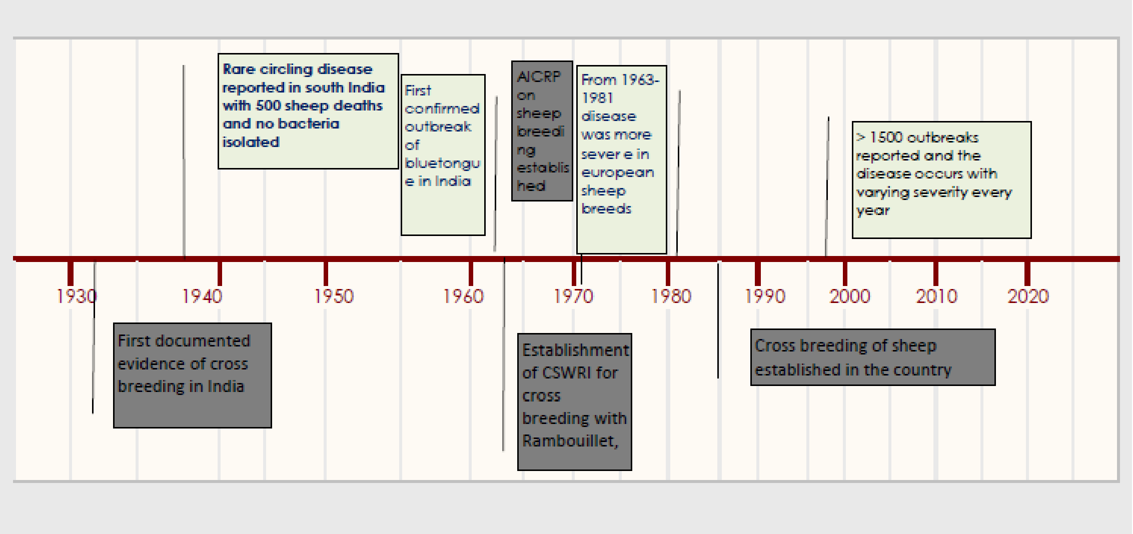

Rare circling disease reported in south India with 500 sheep deaths and no bacteria isolated
First confirmed outbreak of bluetongu e in India
AICRP on sheep breedi ng establis hed
From 1963-1981 disease was more sever e in european sheep breeds
> 1500 outbreaks reported and the disease occurs with varying severity every year
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020
First documented evidence of cross breeding in India
Establishment of CSWRI for cross breeding with Rambouillet,
Cross breeding of sheep established in the country

Figure 1.3: Time line of the bluetongue outbreaks in the past, and crossbreeding of local breeds with exotic sheep in India.

1.4.2 Herd immunity

Immunity against bluetongue virus is mainly elicited by the virus protein 2 (VP2), which also determines the serotype. There is very little cross-protection and it is mainly restricted to serotypes which have similar nucleotide sequence in the VP2 region. Immunity against any one bluetongue serotype may not last more than three years, and there are reports of protection against serotypes which fall within a particular serogroup (a group of serotypes which elicits cross-reacting antibody response) (Maan et al., 2009)

1.4.3 Serotypes

Out of 26 known serotypes, twenty two are circulating in South India, based on virus isolation or antibodies against the serotypes (tested using the serum neutralization test against particular serotypes) (Sreenivasulu et al., 2003). Bluetongue serotypes 1, 2,3,5,6 and 10 are of high pathogenicity, with the potential for causing epidemics in Africa (Dungu et al., 2004). Particular strains within a single serotype can be more pathogenic than others (Saegerman et al., 2007).

1.4.4 Breed and host susceptibility

There are more than 40 breeds of sheep in India of which 14 are present in South India. Sheep breeds can be categorized into four different types; non-descript sheep, local breeds, purely exotic breeds and crossbreeds. These descriptions are based on phenotypic and genotypic characterization of indigenous breeds. Non-descript sheep are defined as indigenous breeds which cannot be identified or do not have more than 50% similarity (phenotypically and genotypically) to any recognized local breed. The other indigenous breeds are considered under the local breed category, with distinct genotypic and phenotypic characters. Crossbreeds of sheep are defined as those breeds which are a mix of exotic and local breeds and those which are mix of just local breeds do not fall under this category.

Fundamental differences have recently been identified in the inherent susceptibility of endothelial cells from cattle and sheep to BTV infection (DeMaula et al., 2001, 2002a 2002b). Pure local breeds of Asia are thought to be resistant compared to the exotic breeds of Europe and America, as demonstrated by the presence of antibodies against BTV in the local breeds of South East Asian countries such as India (Prasad et al., 2009), Indonesia and Malaysia (Hassan et al., 1992 and Sendow et al., 1991) without any clinical signs of the disease.

1.4.5 Bluetongue virus serotypes variability and re-assortment

Bluetongue is a segmented orbivirus which has the ability to "re-assort" or exchange genomic segments with other BTV strains co-infecting the same host. This re-assortment generates the additional threat of fundamental shifts in virulence and in the transmission potential of strains (Shaw et al., 2013). For example, field and vaccine strains of BTV are known to have re-assorted to produce unwanted phenotypes during the recent European epidemic (Batten et al., 2008). Recent reverse genetics research suggests that the re-assortment process is very flexible (it can involve any segment) and is frequent in BTV (Shaw et al., 2013). Worldwide, the impacts of bluetongue viruses have changed in recent decades, with unprecedented outbreaks of multiple BTV serotypes across Europe since 1998 (Saegerman et al., 2008).

BTV serotypes in India are from both eastern and western topotypes (Maan et al., 2012, Maan et al., 2012) and there is a possibility of new re-assortants circulating in the region, as was recently observed in the case of serotype

3 (eastern and western reassorted strain) (Maan et al., 2012). The presence of both the eastern and western topotypes and vaccine strains in India may be due to the importation of livestock from other countries (Rao et al., 2012).

1.4.6 Stress of other diseases, lambing and other stressors

Although bluetongue is the major infectious disease of sheep in India, other diseases are also present. Enterotoxaemia is now controlled by extensive vaccination coverage (~90%), resulting in a drastic reduction in the number of reported outbreaks in the past decade (Ahuja et al., 2008). There is high sero-prevalence, but few clinical outbreaks in sheep, of PPR (Singh et al., 2004) which also occurs as mixed infections with BTV (Mondal et al., 2009). Ecto and endoparasites of sheep are also present, fascioliasis being the most common. Infection with other diseases can make the animals more susceptible to bluetongue which can be more severe in immuno-compromised animals (Brodie et al., 1998). Nutritional deficiencies and the stress of lambing (Erasmus, 1975) can also make animals more susceptible to bluetongue and other viral diseases.

1.5 Extrinsic factors

Apart from the spatial variation in risk of BT in South India, there is temporal variation with impacts varying widely between years. Monsoon conditions like the South-West monsoon and the North-East monsoon are expected to govern within-year seasonality (Prasad et al., 2009), whereas inter-annual variability may be due to the influence of waning herd immunity (intrinsic factor) coupled with stock replacement (biotic extrinsic factor) or inter-annual climate variability (abiotic extrinsic factor).

Among extrinsic factors, temperature plays a role in virus infection of the vectors and also the transmission of the virus to their host (Wellby et al., 1996). Replication of the virus in the midges does not occur at temperatures at or below 15^0C (Carpenter et al., 2011). As the temperature increases, infection rates also increase, viral replication is faster and transmission occurs sooner, although midges survive for relatively shorter times. Seasonality has been observed to have significant influence on the abundance of different *Culicoides* spp (Sanders et al., 2011; Searle et al., 2013).

1.6 How can statistical models help in understanding intrinsic and extrinsic drivers of disease at different spatial and temporal scales?

The methods used in the analyses of outbreak data depend on the spatial and temporal scale of the data and whether the dependent variable is normally distributed, or count data, or presence and absence data. The methods applied in the analysis of epidemiological data will be broadly discussed under presence and absence methods, spatial methods, temporal methods and spatio-temporal methods for count data.

1.6.1 Presence and absence methods

In a classical linear regression model (Eq1), the dependent variable is assumed to be continuous, normally distributed and a linear function of a set of independent variables which may be continuous, categorical or a combination of the two. Multiple regression, analysis of variance and analysis of covariance come under the linear regression model and the parameters are estimated using ordinary least squares method.

$$Y = \beta 0 + \beta 1x1 + \beta 2x2 + \beta nxn + \varepsilon \qquad (1)$$
$$\varepsilon \sim iid\ N(0, \sigma^2)$$

Where β_0 is the intercept, $\beta_{1\ldots}$ β_n are the co-efficients for independent variables and $x_1 \ldots x_n$ are the independent variables, and errors, ε, are normally distributed. These notations will be used in the rest of this chapter, and any additional parameters will be defined wherever required.

Linear regression methods are not suitable for non-normal continuous data, categorical data or count data. Generalized linear models (GLM, Nelder & Baker, 1972) were introduced to overcome these limitations of classical linear regression.

$$\Pr\,(\mathrm{Y} = \mathrm{y}/p) = Binomial(p)$$
$$\mathrm{logit}(\,p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_n x_n \qquad (2)$$

There are two main components in GLM; the link function and an error distribution. The link function is used to transform the mean of the dependent variables. The transformed variable is a linear function of regression parameters. The regression parameters and standard errors are estimated by specifying error probability distribution. Logistic regression is the most commonly employed GLM for binary outcome data. In logistic regression a logit link is used and a binomial distribution is assumed.

Machine learning (ML) methods such as neural networks and support vector machines have been developed for modelling binary outcome data. These methods are less frequently used in ecology or epidemiology due to the difficulty of biological interpretation (Elith & Graham, 2009).

Models for *Culicoides* abundance or for the presence and absence of bluetongue have been developed mostly for Europe. Preliminary spatial models for Bluetongue in Europe were aimed at delineating the areas where adult *C.imicola* could survive year round, or areas where overwintering of BTV was possible (Sellers & Mellor, 1993). Logistic regression was used to predict the distribution of *C.imicola* by using historic monthly weather station data (which did not include rainfall) (Wittmann et al., 2001). Satellite imagery was used to predict *C.imicola* presence and abundance in Europe and North Africa using a discriminant analysis approach (Tatem et al., 2003).

1.6.2 Time series analysis

Ordinary least square methods discussed in the earlier section (Eq1) can also be used to model time series data with independent and normally distributed errors by including a time component with uncorrelated errors (Eq 3). Residuals of time series regressions are rarely uncorrelated, however, especially when present events depend upon past events, as they do in the case of infectious diseases (Selvaraju et al., 2013)

$$Yt = \beta 0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_n x_{nt} + \varepsilon_t \qquad (3)$$
$$\varepsilon_t \sim iid\ N(0, \sigma^2)$$

Errors in time series regressions are split into two components, one time dependent (autoregressive) and one not (and assumed normally distributed) (Eq. 4). In combination, therefore, the total error term tends to be heteroscedastic in time series models rather than homoscedastic (the assumption in standard linear regression)

$$Yt = \beta 0 + \beta 1 x_{1t} + \beta 2 x_{2t} + \beta_n x_{nt} + \varepsilon_t \qquad (4)$$
where
$$\varepsilon_t = \phi 1 \varepsilon_{t-1} + \phi 1 \varepsilon_{t-2} + ... + w_t$$
$$w_t \sim iid\ N(0, \sigma^2)$$

In purely autoregressive models the past observations (t-1, t-2, t-3 etc.) are used as dependent variables and the errors are normally distributed (Eq 5). There are examples in other VBD's which incorporate temporal dependency in the time series with Gaussian outcomes (Luz et al., 2008) using the popular Box-Jenkins methodology (Helfenstein, 1986). Autoregressive time series models have been extended to include weather variables (Gharbi et al., 2011; Helfenstein, 1991) and their lags, and can also account for seasonality using seasonal autoregressive models (Zhang et al., 2010).

In vector borne diseases, the observations are not only dependent on past observation, but also influenced by environmental variables. Purely autoregressive models are best suited for data which are known to be drawn from only endogenous processes (Keitt et al., 2002). The role of temperature and/or rainfall on malaria is well established (Clements et al., 2009; Zhang et al., 2008; Zhou et al., 2004) by accounting for temporal autocorrelation of the dependent variable. But, there are very few studies on bluetongue or other vector-borne animal diseases which establishes quantitative links with meteorological variables, accounting for temporal autocorrelation of the dependent variable. In a study to disentangle the roles of temporal dependence and climate (Zhou et al., 2004), it was demonstrated that climate dominates over intrinsic factors (past cases of malaria), but once temporal dependence was accounted for (Singh & Sharma, 2002) the effect of rainfall on malaria became nonsignificant. Therefore failure to account for the temporal autocorrelation (as in malaria cases) can lead to selection of non-significant variables or the elimination of significant variables.

$$y_t = c + \phi 1 y_{t-1} + \phi 2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t \qquad (5)$$

In equation 5, c is constant, the various ϕ s are autoregressive parameters and ε_t is white noise (normally distributed errors).

Autoregressive time series methods are well established for Gaussian outcomes, whereas models for non-Gaussian count data are less developed in environmental epidemiology (Bhaskaran et al., 2013) and are very few in infectious disease epidemiology (Lu et al., 2009). The methods for non-Gaussian data (Eq 6) include GLMs (Generalized Linear Models) and other non-parametric regression methods. GLMs that incorporate temporal dependence as random effects (Eq 7) are referred to as GLMM (Generalized Linear Mixed Models). In GLMs or GLMMs all parameters are estimated as fixed effects and estimated by maximum likelihood methods or quasi likelihood methods (to account for over dispersion).

$$\begin{aligned} &\Pr(\mathrm{Y} = \mathrm{y}/\mu) = Poisson(\mu i) \\ &\log(\mu\mathrm{t}) = \beta 0 + \beta 1x1t + \beta 2x2t + \beta nxnt \quad + \varepsilon t \end{aligned} \tag{6}$$

$$\begin{aligned} &\Pr(\mathrm{Y} = \mathrm{y}/\mu) = Poisson(\mu_i) \\ &\log(\mu_t) = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_n x_{nt} + \varepsilon_t \\ &\varepsilon_t = \phi_1 \varepsilon_{t-1} + \phi_1 \varepsilon_{t-2} + \ldots\ldots + w_t \\ &w_t \sim iid\ N(0, \sigma^2) \end{aligned} \tag{7}$$

1.6.3 Spatial analysis of count data

Linear regression models (Eq 1) can also be used in analysis of spatial data by ignoring the spatial correlation which may lead to bias in estimates. In disease epidemiology, observations which are nearer to each other tend to be more related than observations farther apart. Presence of this spatial autocorrelation inflates model accuracy but also the estimated explanatory power of environmental predictors (Dormann et al., 2007). Spatial autocorrelation can result from intrinsic biotic processes such as disease spread to neighbouring areas but can also be a problem if certain independent abiotic variables (with spatial structure) are omitted or not measured. Ignoring spatial autocorrelation can also lead to undue emphasising of non-significant relationships between the response variable and environmental variables.

Spatial autocorrelation in epidemiological data

Disease data are most often aggregated at a particular administrative level, and referred to as 'polygon data'. In rare experiments the data are collected on regularly spaced grids and are known as lattice data (Bivand et al., 2008). Neighbours can be defined by contiguity based methods or graph based methods. Contiguity based methods involve a two-step process. First, neighbours are defined based on the adjacency matrix and second the weights are assigned to neighbours with which they share a boundary (Bivand et al., 2008). The adjacency matrix for a graph (a graph is a set of points called vertices and a set of lines called edges) having n vertices is an nxn matrix whose (i, j) entry is 1 if the ith vertex and jth vertex are connected, and 0 if they are not. In graph-based methods the polygon centroids are used to calculate the distance between two polygons. The distance between two centroids is calculated using different triangulation methods (Bivand et al., 2008).

The neighbourhood can also be defined using distance based methods. For example, in in the k-nearest neighbourhood method the k nearest points are considered as neighbours; however this method is most suitable for regularly spaced data and may not be suitable for irregular polygon data (districts) (Bivand et al., 2008) which is more common in epidemiology.

Spatial autocorrelation (SAC) can be detected by examining the residuals of non-spatial models using different methods such as Moran's I, Geary's

C and variograms (Bivand et al., 2008), widely used in geostatistics (Isaaks & Srivastava, 1989). Moran's I values show the presence and absence of SAC and range from -1 to 1 as follows; zero is complete independence, or absence of SAC; 1 is dependence or presence of positive SAC i.e. residual error values are more similar for observations that are close together; -1 is presence of negative SAC where residual error values are less similar for observations that are close together. In epidemiological analyses, only the presence of positive SAC is plausible. Moran's I based correlograms plot Moran's I values against distance for different distance bins. As the distance increases the spatial autocorrelation decreases, reaching a value of zero where the observations are completely independent. The values of Moran's I can be tested for significance using statistical tests. Variogram plots are the inverse of Moran's I correlogram.

Methods to account for spatial autocorrelation in count data

Spatial autocorrelation in the residuals of non-spatial models can be accounted for using different methods (Dormann et al., 2007). The Generalised Least Squares approach (GLS) (Venables & Ripley, 2002), which is a modification of the Ordinary Least squares (OLS) method, can be applied when the errors are correlated. The errors can be treated with different correlation structures depending on the definition of neighbours. If the adjacency matrix is used for defining neighbours then the parameters can be estimated by GLS using a CAR (conditional autoregressive) (Keitt et al., 2002) or SAR (simultaneous autoregressive) model, equations 8 and 9 respectively (Haining, 2003), which differ in whether or not they can handle the asymmetric covariance matrix. The spatial neighbourhood matrix contains spatial weights (W) and this matrix is symmetric (with zeroes on the diagonal and weights for the neighbouring locations on the off-diagonal which can be identical even after transposing) in a CAR model and the neighbourhood matrix can be asymmetric (with zeroes on the diagonal and spatial weights on the off-diagonal which will not be identical after transposing) in a SAR model. There are three methods by which a SAR model can be specified. In the first method, SAC in the response variable is accounted for using the spatial proximity matrix of the observations. In the second method, SAC in the response and the predictor variable is accounted for in the model. The response variable is not only dependent on the magnitude of covariates in the areas, but also on the covariates in the neighbouring areas. In the third method, the errors are assumed to be dependent only on the neighbours and not on either the response or predictor variables.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_n x_n + \rho C(Y - \beta_n x_n) + \varepsilon \qquad (8)$$
$$\varepsilon = iid\ N(0, \sigma^2)$$
$$\rho = \text{autoregressive parameter}$$
$$C = \text{symmetric neighbourhood matrix}$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_n x_n + \rho W(Y - \beta_n x_n) + \varepsilon \qquad (9)$$
$$\varepsilon = iid\ N(0, \sigma^2)$$
$$\rho = \text{autoregressive parameter}$$
$$W = \text{asymmetric neighbourhood matrix}$$

If distance based measures are used to define neighbours (geostatistics) then different correlation structures such as exponential, Gaussian or spherical can be specified (based on the semi-variogram) and estimated in a GLS framework (Eq 10). In GLS the response variable is assumed to be normally distributed with a specific correlation structure.

$$
\begin{aligned}
&Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_n x_n + \varepsilon \\
&\varepsilon = N(0, \Sigma) \\
&\Sigma = \text{correlation structure} \\
&\text{(Gaussian, exponential, spherical)}
\end{aligned}
\qquad (10)
$$

Non-normal data can be modelled either by a Generalised Linear Model (GLM) (Eq 11) or by GLMM (Generalised Linear Mixed Model) with correlated errors (Eq 12).

$$
\begin{aligned}
&\Pr(\mathrm{Y} = \mathrm{y}/\mu) = Poisson(\mu_i) \\
&\log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_n x_{ni}
\end{aligned}
\qquad (11)
$$

$$
\begin{aligned}
&\Pr(\mathrm{Y} = \mathrm{y}/\mu) = Poisson(\mu_i) \\
&\log(\mu_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_n x_n + \varepsilon \\
&\varepsilon = CAR + w \\
&w \sim iid\ N(0, \sigma^2)
\end{aligned}
\qquad (12)
$$

In auto covariate based GLM, an additional, distance-weighted covariate is included in the model, and the weights are assigned based on neighbours (Kaboli et al., 2006). The additional covariate (auto covariate) is calculated based on the influence of neighbouring values on each particular location. Spatial Eigen Vector Mapping (SEVM) is also a modification of GLM, and here there is decomposition of the connectivity matrix into different Eigen vectors. Those Eigen-vectors which reduce the spatial autocorrelation are used as predictors in a GLM model. In ecological studies, when there is spatial dependence, and key covariates may be missing, the ML approach often leads to unsatisfactory estimates of the district level risk due to extra-Poisson variation.

1.6.4 Space-time analysis of epidemiological data

Analysis of space-time data in epidemiology has benefitted from advances in both modelling and geographical information systems. Selection of the particular analytical method to use in any situation is also important and depends on a variety of factors such as the spatial and temporal scales of the data (spatial; local/regional/national and temporal; day/week/month/year), on data quality and data type (cases, outbreaks, mortality) (Robertson et al., 2010).

Space-time analysis was initially aimed at testing for the presence of space-time interaction, or the presence of clusters. Detections of clusters in both space and time simultaneously are extensions of spatial cluster detection methods (Besag & Newell, 1991). Interactions are said to occur when, for example, disease cases occur closer together in either or both of space and time than would be the situation without interaction, when cases would occur at random in both space and time. The demonstration of space-time interaction in infectious diseases helped to understand the spread of such diseases into neighbouring areas (Waller et al., 2007). In the case of non-infectious diseases, it can help to identify any underlying geographical cause (for example, a restricted area of rare gas emission, or a local concentration of a chemical contaminant).

Tests for the presence of space-time interaction using the null hypothesis of no-interaction fall into one of three types (Robertson et al., 2010); i) tests for space-time interaction, ii) Cumulative Sum (CUSUM) methods, and iii) scan statistics. Space-time interaction tests include the Knox test (Paré et al., 1996), the Mantel statistic (Ward & Carpenter, 2000) or their modifications. These statistical tests require individual case data (i.e. the location and timing of each individual infection). In the Cumulative sum (CUSUM) method, the presence of an interaction is detected by a cumulative alarm statistic, which indicates change in an underlying process over a period of time. Scan statistics (Kulldorff et al., 2007) are another class of method to test space-time interaction and often used to detect outbreaks in space and time.

The methods discussed so far (clustering, space-time interaction) are not designed to quantify the relationship between the dependent and independent variables and hence may not be suited for making predictions in unknown areas. The methods discussed in the spatial and temporal domain (linear, GLM or GLMM) can be extended in the space-time domain and used for making predictions in unknown areas. Spatial and temporal autocorrelation parameters can be modelled as fixed parameters using maximum likelihood estimation in the frequentist domain or as random effects in the Bayesian domain (Clayton, 1996).

1.6.5 Bayesian methods

Bayesian methods differ from maximum likelihood methods by inclusion of prior information about the parameters. If the data are highly informative with a weak prior, then similar answers are generated as in the maximum likelihood approach (Bolker et al., 2009). In the frequentist approach, the parameters such as the mean, variance and regression coefficients are fixed but unknown, and are calculated from the data. In Bayesian methods, the parameters are not fixed but follow a statistical distribution. The Bayesian way of estimating parameters can be demonstrated by giving the Bayesian equivalent of equation 1 for simple linear regression.

$$Y = \beta_0 + \beta_1 x_{1} + \beta_2 x_2 + \beta_n x_n + \varepsilon \qquad (13)$$

$$\varepsilon \sim N(0,\tau)$$

$$\theta = (\beta,\tau) \text{ parameters}$$

$$\text{f}(\theta) = \text{prior for } \theta$$

$$\text{or}$$

$$\beta \sim N(\mu_\beta,\sigma^2)$$

$$\tau \sim gamma(a,b)$$

$$\text{Then the likelihood can be specified as}$$

$$f(\theta \mid Y) = f(Y \mid \theta) \times f(\theta / f(\text{Y}) \propto f(y/\theta) f(\theta)$$

Y are the observed data and β the regression parameters. The errors are normally distributed with mean zero and variance (τ). The unknown parameters (θ) are drawn from a statistical distribution. We can assign informative priors for the parameters based on expert opinion or prior experiments or we can specify non-informative priors if there is no pre-existing information about these parameters. The complexity increases as the number of random effects increases, as in the GLMM models, and parameter estimation using maximum likelihood methods may lead to bias in estimates. Bayesian inference can then be advantageous over frequentist methods.

Although Bayesian methods have been well developed in spatial mapping of human diseases (Best et al., 2005; Lawson, 2013; Sanders et al., 2011), there are very few examples to date in veterinary epidemiology (Schrödle & Held, 2011). In the current context *Culicoides* distribution data have been analysed using Bayesian methods (Sanders et al., 2011; Searle et al., 2013).

1.7 Aims and structure of the thesis

The aim of this thesis is to understand the role of intrinsic and extrinsic factors in determining the severity of bluetongue outbreaks across South India using statistical models and to determine whether these models can contribute towards development of an early warning system for the disease at different spatial and temporal scales.

Chapter 2 deals with the application of remotely sensed variables to discriminating presence and absence sites of bluetongue outbreaks using non-linear discriminant analysis. The use of known absence data in endemic areas is highlighted in this chapter, and model results using them are contrasted with those of models that use only pseudo-absence data. In chapter 3, attempts are made to disentangle the role of intrinsic and extrinsic factors in driving annual and inter-annual variability of bluetongue outbreaks in Andhra Pradesh and to develop a temporal forecasting model for making future predictions. In chapter 4 the influence of different hosts and breeds, land cover and climate on the severity of bluetongue outbreaks at district level is analysed, taking into account spatial autocorrelation. Chapter 5 deals with the roles of spatial and temporal heterogeneity and climate in driving inter-annual variability of bluetongue outbreaks at the district level. In chapter 6, Bayesian Network Modelling (BNM) is employed to identify the direct and indirect associates of bluetongue cases at the village level with host, land cover and climate variables. The variables identified in the BNM model are then used to develop a predictive model of BTV severity in unknown areas. Finally chapter 7 concludes by discussing the implications of the study for bluetongue control and the potential contribution of the models described here to the development of early warning systems for BTV.

Chapter 2

A Non-Linear Discriminant Analysis approach to distinguish areas of presence and absence of Bluetongue in South India

2.1 Introduction

Predictions of the presence and absence of Bluetongue outbreaks using logistic regression have been, and continue to be, employed at the district level in India (www.nadres.in). There are more than eighty thousand villages in three states of South India (Andhra Pradesh, Karnataka and Tamil Nadu) and villages in close proximity can differ substantially in the severity of bluetongue outbreaks. Understanding the environmental conditions suitable for presence and absence of bluetongue at the village level is necessary for effective control and future surveillance. In this chapter temporal Fourier processed remotely sensed variables are used to produce risk maps for bluetongue using a Non-Linear Discriminant Analysis approach (NLDA).

Understanding the relationship between presence and absence of disease or its vectors and environmental variables is important in defining risk areas and making predictions in unknown areas (Rogers and Randolph, 2003; Palaniyandi 2012). In the past, weather station data have been used for spatial and temporal predictions of many vector borne diseases including bluetongue (Calistri et al., 2003; Conte et al., 2003; Wittmann et al., 2001). Weather station data have their own advantages and limitations for such work. The weather station data have coarser spatial resolution, but certain parameters like air temperature, air humidity and precipitation that have direct impact on the life history of different *Culicoides* species can be measured, though this is rarely done in the microhabitats actually used by adult insects (Purse et al. 2004). More-over, the presence of only a few weather monitoring stations can, however, hamper the quantification of relationships between the vector and weather station environmental variables.

A logistic regression model for *C.imicola* in the Iberian Peninsula using averaged climate data was extrapolated across the entire Mediterranean Basin by (Wittmann et al., 2001). Conte et al., 2003 also used averaged climate data and a logistic regression approach to study the effect of climate on the presence of *C.imicola* in Italy. They obtained an overall 75% correct classification but there was evidence of some misclassification due to geographical clustering.

Remotely sensed variables have been used as environmental variables or surrogates of meteorological variables in spatial and temporal models of many vector borne diseases (Kalluri et al., 2007; Kitron1998; Kitron, 2000; Rogers & Packer 1993; Rogers et al., 2002). Many such data are free and obviously cover much wider geographical areas in more detail than do meteorological station records. There are many studies reporting NDVI (Normalized Difference Vegetation Index) as one of the important predictors in the models of bluetongue vector distribution and abundance (Baylis et al., 1998; Tatem et al., 2003). However, when the models were used to predict the distributions of different vector species using remote sensed variables, different sets of variables were selected as there are differences in the life history requirements of different *Culicoides* species (Purse, Tatem, et al., 2004; Purse et al. 2012).

Presence and absence of a disease or vector is not only determined by the average values of environmental variables (remotely sensed or weather station data), but also often by their seasonality. Seasonality is just as pronounced in tropical and subtropical countries such as India as it is in temperate and sub-temperate regions, although the obviously seasonal variables tend to differ (rainfall in the tropics and temperature in temperate regions). Use of raw time series data (monthly remotely sensed variables or weather station data) is not advised because of serial correlation in the data. Principal component analysis (PCA) (Eastman & Filk, 1993) is the most common technique in data ordination methods, but seasonality is lost (Hay et al., 1998; Rogers et al., 1996). Temporal Fourier analysis (TFA) of remotely sensed variables (Scharlemann et al., 2008) overcomes the problem of serial correlation and also helps to capture the seasonality in the environmental conditions. The results of TFA of any single data channel are no longer serially correlated and can be used as independent predictors or discriminating variables (although the TFA products of related channels, for example day- and night-time Land Surface Temperature, may be correlated). The use of TFA imagery and their advantages have been discussed in mapping different vector borne diseases (Rogers et al., 1996) and bluetongue (Tatem et al., 2003).

There are many *Culicoides* species reported from South India, of which *C.imicola* (Mellor et al., 2000), *C.brevitarsis* (Muller et al., 1982) and *C.oxystoma* (Mellor et al., 2000) are proven vectors of the bluetongue virus. The habitat requirement for each species varies and in the absence of species' distribution data it is difficult to understand the role of different environmental variables in determining their abundance which ultimately causes outbreaks in South India. Most (60 out of 80) of the constituent districts of South India have reported bluetongue outbreaks and, among them, the villages in each district may not be equally favouring the transmission because different species may be involved, each requiring different environmental conditions for breeding and subsequent virus transmission. Thus, in endemic countries in which different species of *Culicoides* are involved in transmission and different environmental factors play a role, there is a possibility of the presence of more than one epizootiological system. Therefore, it is important to understand the role of environmental variables in discriminating between the different systems. In this chapter, a Non-Linear Discriminant Analysis (NLDA) is applied to bluetongue presence and absence data in three South Indian states with the following objectives:

1. To identify the variables which discriminate between presence and absence of bluetongue outbreaks in South India.
2. To investigate whether different environmental conditions operate in South India to discriminate between presence and absence of bluetongue outbreaks (more than one presence and absence groups)?
3. To test whether known but sparse absence data or pseudo-absence data give better accuracy in presence/absence models.
4. To develop a BTV risk map for South India to help in future surveillance.

2.2 Materials and methods

2.2.1 Bluetongue presence and absence data

Bluetongue presence and absence data were collected from three states with different reporting systems. Data for Karnataka were obtained from the State Animal Disease Monitoring and Surveillance, Bangalore and also from NIVEDI (National Institute of Veterinary Epidemiology and Disease Informatics). Records in the NIVEDI database are only of Polymerase Chain Reaction (PCR)-confirmed BTV virus presence in field-collected samples. The data for Andhra Pradesh were obtained from the State Department of Animal Husbandry, Hyderabad. Data for Tamil Nadu were obtained by visiting different districts which were known from historical records to be endemic (Tirunelveli, Madurai, Karur, Dindigul, and Erode). For other districts in Tamil Nadu, the data were collected from the Central Referral laboratory (CRL), Chennai. There was no information on the diagnostics tests performed from the records of Tamil Nadu (both from CRL and during visits to districts). All presence records for bluetongue were obtained for the years 1997 to 2011. No village level recording could be found before 1997. In the past there was lack of sensitive diagnostics (like PCR) for detection of BTV virus or antibodies (competitive ELISA).

The records for Karnataka and Tamil Nadu contained only names of disease-affected villages whilst the records for Andhra Pradesh had the number of cases in each village. Only simple presence and absence records were considered in this analysis for all the three states, but case data for Andhra Pradesh was analysed as discussed in Chapter 6. The village names were first matched with the village databases for all the three states and the village centroids were then extracted for all the presence and absence records from the village level shape files (obtained from the Survey of India through an individual license). There were in total 769 villages in all the three states of South India that recorded BTV presence at some point in time between 1997 and 2011from these data sources and 59809 villages not reporting bluetongue at all. A village might have no records of BTV outbreaks either because it is genuinely disease-free (because conditions there are unsuitable for BTV; a genuine absence site) or because the village is environmentally suitable for BTV but has never experienced any outbreaks purely by chance (a potential presence site, recorded as absence), or because outbreaks have occurred there but have not been reported for various reasons (a genuine presence site, but recorded in the database as absence). Because there is a variety of reasons for the recording of absence of BTV in the village database, the presence/absence models were run twice, once using the database-recorded village absence sites and once using ‘pseudo-absence’ data generated by a relatively standardised method in the eRiskMapper software employed. Pseudo-absence points are generated by many presence/absence packages because most databases of animal or plant species or diseases fail to record genuine absences (records are often based on data from museum collections of specimens); hence the absence data must be generated in one way or another. In eRiskMapper the user is allowed to define both a minimum and a maximum distance from any presence site where pseudo-absence sites might be randomly selected. The minimum distance attempts to guarantee that no pseudo-absence site is so similar in environmental conditions to a genuine presence site that it might actually be a suitable site for the species under study. The maximum distance attempts to guarantee that pseudo-absence sites are not so different from any presence site that their data are unhelpful in distinguishing local presence and absence (for example it would be inappropriate to include an Arctic pseudo-absence site in a database for a tropical disease). In the present models the minimum distance was set to 0.5 degrees of latitude/longitude (approximately 50-60 km at the equator) and the maximum to ten degrees. However, the mask used to select pseudo absence sites were restricted to the three South India states and the adjoining states (Kerala, some parts of Maharashtra and Orissa).

2.2.2 Remotely sensed variables

In total, 50 temporal Fourier processed MODIS variables (at 1km spatial resolution) were used in the analysis (Scharlemann et al., 2008). The temporal Fourier processing extracts the seasonal information of the remotely sensed variables and describes it in terms of the mean, the annual minimum and annual maximum, the amplitudes and phases of the annual (a1 and p1 respectively), bi-annual (a2, p2) and tri-annual (a3,p3) components of the signal and finally the variance. The MODIS channels processed in this way were the MIR (Middle Infra-Red), daytime Land Surface Temperature (dLST), night time Land Surface Temperature (nLST), NDVI (Normalized Difference Vegetation Index) and the EVI (Enhanced Vegetation Index). The NDVI is a measure of photo synthetically active radiation (PAR) and has been variously interpreted as an indicator, directly or indirectly of chlorophyll abundance, vegetation biomass, soil moisture and rainfall (Campbell, 2002). MIR is correlated with water content, surface temperature and the structure of vegetation canopies, especially in young forest re-growth stands (Boyd & Curran, 1998).

2.2.3 Non-Linear Discriminant Analysis Model (NLDA) description

Non-linear discriminant analysis was carried out using a Windows based eRiskMapper package (David Morley, Luigi Sedda and David J Rogers, 2011), based on the previous (non-Windows-based) software of Rogers (Rogers 1993; Rogers & Randolph 1993).

Discriminant analysis is rather different from a number of other methods used to describe species distribution because it is a classification-based rather than regression-based technique. The algorithm assigns each observation to one or other category in a mutually exclusive set of groups that encompasses the entire range of variability expected. One classical application of discriminant analysis is the assignment to one or other of several hominoid lines of fossils of a newly discovered human-like skull or other body part. The assumption made is that the new fossil must belong to one or other group and that all the groups considered in the analysis comprise the entire fossil history of humanoid apes (i.e. there are no missing groups in the pre-existing recorded fossil record). If the new fossil cannot be assigned to any existing group with any certainty then it is often referred to as a ‘missing link’, and a new group (based on a sample of one) is defined, to contain it.

The assumptions of linear discriminant analysis, the simplest form of this technique, are as follows:

1. There must be a minimum of two or groups.
2. The number of observations in each group can be a minimum of two but in practice should be a minimum of at least 30, to define each group sufficiently precisely (this minimum increases with the number of variables used to define each group).
3. The number of discriminating variables should be no more than two fewer than the number of observations (n-2, where is n number of observations). Again in practice the number of observations should exceed the number of variables by a much bigger margin than this, since a model with almost as many variables as observations will be near perfect (since each variable can capture one of the observations) but will have no residual degrees of freedom for any significance testing.
4. The covariance matrices for each group (e.g. presence and absence) are the same (there is an important difference between LDA and NLDA in this respect, discussed later).
5. Each group has been drawn from a population with a multivariate normal distribution of all discriminating variables, which are usually continuous (dummy-coded categorical variables may also be used with care).

The main aim of discriminant analysis is to use the covariance matrix to calculate a discriminant function to assign group membership to which the new observations will be assigned.

The discriminating variables are used to predict the group to which the particular observation belongs based on the location of the observation with respect to the group centroid. The Mahalanobis distance (Eq. 1) is commonly used to classify group membership and was used in this study. The Mahalanobis distance, MD, (D^2 in equation 1) is a covariance adjusted measure of difference between sets of environmental conditions in this application; small values indicate similar conditions and large values dissimilar ones (Rogers 2015). It is calculated as follows:

$$\begin{aligned} D^2{}_{12} &= (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)'\mathbf{C}_W{}^{-1}(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2) \\ &= d'\mathbf{C}_W{}^{-1}\mathbf{d} \end{aligned} \qquad (1)$$

The subscripts refer to group 1 (for BTV presence) and 2 (for BTV absence), $\overline{\mathbf{X}}_1$ and $\overline{\mathbf{X}}_2$ are the mathematical vectors of the mean values of the discriminating variables defining each group (i.e. their centroids), d= $(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)$ and $\mathbf{C}_W{}^{-1}$ is the inverse of the within-groups covariance (dispersion) matrix. Once D^2 is calculated between any observation (where the mean vector in equation 1 is replaced by the vector of values for that particular point) and the centroids of each group in turn, an observation is classified as belonging to that the group to which it has the smallest D^2. The group with smallest D^2 is the one in which the environmental conditions most closely resembles the profile of this observation. If D^2 is large then the profile of this observation may match poorly to this group, but may be better than its match to any other group.

We assume that each group comes from a population with a multivariate normal (MVN) distribution. Most of the observations will be clustered near the centroid and the number of observations will be less as we move away from the centroid. By knowing the distance from the centroid, we can know the proportion of the group's population that is closer and the proportion that is further away. The proportion of the group's population that is further away from the centroid is the probability that an observation located that far away actually belongs to the group. Classification of an observation into the closest group according to D^2 assigns it to the group to which it has the highest probability of belonging.

The D^2 may be turned into the posterior probability of belonging to the different groups in the analysis. This is achieved effectively by inserting the MD into the equation for the standard normal distribution of which the clusters are samples, using the following equations

$$\Pr(1/x) = \frac{p_1 e^{-D^2_1/2}}{\sum_{g=1}^{2} p_g e^{-D^2_g/2}} \qquad (2)$$

$$\Pr(2/x) = \frac{p_2 e^{-D^2_2/2}}{\sum_{g=1}^{2} p_g e^{-D^2_g/2}}$$

Where $\Pr(1/x)$ is the posterior probability that observation x belongs to group 1 and $\Pr(2/x)$ is the posterior probability that it belongs to group 2. The exponential terms in the equation 2 are those of the multivariate normal distribution defining groups 1 and 2; all other terms of the multivariate distributions are the same in the numerator and denominator and therefore cancel out.

p_1 and p_2 are the 'prior probabilities', that is, the probabilities with which any observation might belong to either group given prior knowledge. If there is no prior information about the observation then it is common to assume equal prior probabilities. The assumption of equal prior probabilities is ensured by selecting equal numbers of presence and absence observations in the models (selection of bootstrap samples).

NLDA and Clustering

One of the assumptions of discriminant analysis is that all groups have the same covariance. This is usually violated in species' distribution examples and it is necessary to adapt the equations to allow for different covariance of the different groups (presence and absence in the simplest case). This is shown in Eq 3. Where C_i is the group-specific rather than common covariance matrix, and D_i the corresponding MD. The line of equal probability between the presence and absence groups in multi-variate space is now no longer linear, and so the technique is now described as non-linear discriminant analysis (NLDA).

$$\Pr(1/x) = \frac{p_1 |\mathbf{C}_1|^{-1/2} e^{-D^2_1/2}}{\sum_{g=1}^{2} p_g |\mathbf{C}_g|^{-1/2} e^{-D^2_g/2}} \qquad (3)$$

$$\Pr(2/x) = \frac{p_2 |\mathbf{C}_2|^{-1/2} e^{-D^2_2/2}}{\sum_{g=1}^{2} p_g |\mathbf{C}_g|^{-1/2} e^{-D^2_g/2}}$$

Where $\left|\mathbf{C}_1\right|$ and $\left|\mathbf{C}_2\right|$ are the determinants of the covariance matrices for groups g=1 and 2, respectively (Tatsuoka & Lohnes 1988).

Another commonly encountered problem is that either presence or absence cluster (or both) is not multi-variate normal. To overcome this problem the user of eRiskMapper is allowed to divide the observations into a number of clusters each for presence and absence using the *k-means* cluster algorithm. This algorithm essentially makes a series of clusters each for presence and absence observations, each of which is closer to multi-variate normality than the original group from which those observations were drawn. Thus clustering allows the requirement of NLDA to be met, of a MVN distribution for each class to which observations may be assigned. The number of groups (g in equation 3) then increases to the total number of clusters (presence + absence) selected.

Variable selection

Identification of variables which are important in discriminating presence and absence of any disease is a critical step in NLDA. eRiskMapper uses a forward step-wise procedure whereby variables are included one at a time on the basis of their ability to improve discrimination of the groups within the dataset above that of any other variable not yet included in the selection procedure. The discriminating criteria may be selected from a small group by the user, and these criteria include AIC, AIC$_c$, F-test, Mahalanobis distance, AUC, kappa (details of which are explained later). eRiskMapper automatically chooses up to ten variables in each model although not all ten need be used in fitting the data (for example the AICc criterion may suggest a smaller number, to avoid over-fitting)

Variable selection criteria

In this study the AICc (corrected Akaike Information Criteria) (Hurvich & Tsai, 1989) was used as the selection criteria for including a variable by the forward step-wise method, for eventual use in modelling. If inclusion of an additional variable decreases the AICc by more than a threshold value (5 AIC_C units) then the variable is included; if less then it is not.

Bootstrapping

Sparse datasets are very common in species' and diseases' distribution modelling. Bootstrapping is one of the methods which can be employed on sparse data. It examines the likely importance of variability within the training set on overall model predictions. Thus, for example, if several bootstrapped models give very different results, the training set itself must be very variable and it is therefore highly unlikely that it captures the full range of conditions in which the species occurs in nature. If, on the other hand, the bootstrapped models give more or less similar answers, then it is likely that the training set represents the full range of conditions in nature. Thus it is imagined that the relationship between reality and the training set is the same as that between the training set and its bootstrapped samples. We do not know *a priori* the nature of the first relationship here, but we can investigate it by exploring the second relationship. In this study, 100 bootstraps were generated and predictions were made for each bootstrap and combined to get the average prediction, the final risk map. In each bootstrap 200 points were selected for presence and 200 points for absence groups. The arrangement of samples to have equal number of presence and absence observations in each bootstrap produces model outputs with greatest accuracy (McPherson et al., 2004)

Accuracy and Validation statistics

Different accuracy statistics such as sensitivity & specificity (Congalton, 1991, Fielding and Bell, 1997) or Kappa (Landis & Koch, 1977; Robinson, 2000; Rogers, 2006), were computed for each bootstrap model and the average accuracy statistics of 100 bootstraps presented. Sensitivity measures the proportion of actual presence sites predicted as presence and specificity measures the proportion of actual absence sites predicted as absence. Kappa (k) is an index of agreement that is often used to assess model accuracy and varies between -1 (complete disagreement between predictions and observations) to 1.0 (complete agreement). Kappa=0 when the predictions are no better than random (thus some sites are predicted correctly). Kappa values of <0.4 indicate poor models, of between 0.4 and 0.75 good models, and of greater than 0.75 excellent models (Landis and Koch 1977).

Other accuracy statistics such as AUC (Area under Curve) and Producer's and Consumer's accuracies were also calculated for the 100 bootstrap models. The percentage of all known sites (both presence and absence) correctly classified by the model gives the Producer's Accuracy and the Consumer's accuracy is the percentage of model predictions (presence and absence) that are correct. The kappa, sensitivity and specificity values were also calculated for the test data (hold out) for each bootstrap and the average validation statistics calculated.

Three different sorts of models were run. In the first model (Model 1 series), the villages with no reported BTV outbreaks during the study period were taken as the absence sites; this model was run with one presence and two absence clusters. In the second (Model 2 series) pseudo absence data were generated and used as described in the Materials & Methods section, again with one presence and two absence clusters. Model 3 series also used pseudo-absence data, and three presence and three absence clusters.

2.3 Results

Average accuracy statistics for the three models (Table 2.1) show that Model 1 was the worst and Model 3 was marginally better than Model 2 (kappa was the same, but sensitivity and specificity were slightly higher).

Similarly the average validation statistics of model 1 is worst among the three models. Model 3's kappa value indicates a slightly worse fit than Model 2 but its sensitivity and specificity again were marginally higher than Model 2's. The following section therefore gives details of the single best bootstrap model from the Model 3 series. It should be emphasised that any single model provides only a 'snapshot' of the information contributing to the overall final risk map, the average of 100 model outputs, each with a different bootstrap sample and each possibly with a different set of predictor variables.

2.3.1 Single best model results

The mean values of the predictor variables in the best single Model 3 bootstrap model is shown in Table 2.2 with the key for its variables in Table 2.3. The presence clusters (P1, P2 & P3) and absence clusters (A1, A2 & A3) are arranged row wise and the columns (1-10) represents the variable number as per their order of selection (an indication of their importance in the model). The last column (11) indicates the sample size of each cluster. Table 2.2 shows that bluetongue occurs in areas with low values of variance and bi-annual amplitude of night time land surface temperature (variance =9.18 *vs* 12.76, bi-annual amplitude = 1.32 *vs* 1.71, bi-annual phase =3.28 *vs* 3.52) than in areas with seasonally variable night time land surface temperature. However, high values of the annual and tri-annual phases of nLST (annual=5.23 *vs* 5.18 and tri-annual phase= 1.25 *vs* 1.13) i.e. later seasonal peaks of night time temperature favour bluetongue transmission. The areas with high NDVI (mean 0.41 *vs* 0.46) and EVI (mean =0.26 *vs* 0.28), reflective of dense forest environments, do not favour bluetongue transmission.

Comparing the roles of different Fourier variables in each presence and absence cluster shows slightly different results when compared to mean values. For example, nLST variance is higher in P2 compared to A2 and A3 (variance = 11.0 in P2 vs 7.5 and 3.0 in A2 & A3 respectively), which is in contrast to the overall means for the presence and absence sites for this variable (overall, nLST variance is lower in presence than in absence sites).

Similarly, the tri-annual phase and annual phase of nLST is lower in P1, P2 and P3 compared to A2 and A3.

The model accuracy matrix (also called a ‘confusion matrix’) for the best model is shown in Table 2.4. The matrix shows the observed and predicted category membership. In one presence cluster (P1), there were 62 observed absences and 52 were assigned to this category correctly and only 2 observations were assigned to the absence category. Likewise, out of a total 200 presence points, 32 were misclassified (only 3 were misclassified in the absence category) and among 200 absence points, only 11 points were misclassified.

The overall accuracy result of this best model is shown in Table 2.5. The Producer’s and Consumer’s accuracy are very high (>80%) for all the categories except for the presence 3 category which had 67.4 % Consumer’s Accuracy. In such a Table some misclassification errors are more serious than others. For example, from the Consumer’s point of view an absence site belonging to one cluster but assigned to another absence cluster is a less serious error (since both clusters refer to absence) than one where the site is assigned to a presence cluster. If we combine all presence and all absence clusters before calculating these statistics, the overall Consumer’s accuracy for this model is 96.6% for predicted presence sites and 98.5% for predicted absence sites. In other words, a user of this map can have fairly high confidence in the predictions it has made.

2.3.2 Averaged bootstrap model results

The final risk maps for the three different models (in each case the average of 100 bootstrapped sample models) are shown in Fig. 2.3 and the top ten variables are presented in Table 2.6. The rainbow plots (Rogers, 2006) showing the relative importance of all predictors in all 100 models are shown in Fig. 2.2. The top ranked variable in the Model 1 series is the maximum night time land surface temperature and in the Model 2 series is the minimum night time land surface temperature. In Model 3 the top variable was the tri-annual phase of night time land surface temperature. Except for one NDVI variable each in Model 1 and Model 3, temperature variables dominated in all the model series.

The final risk map for the Model 1 series (Fig. 2.3) is strikingly different from those of the other two models, and indicates large areas of high risk mainly in Andhra Pradesh. The Model 2 and Model 3 series of models, that were much more accurate in validation, predict larger areas of high risk across all three states, but correctly predict low risks in the Western areas (Western ghat forest region), Northern parts of Karnataka, North and a few North-Eastern and Central parts of Andhra Pradesh (Eastern ghat forest regions).

The distribution of the three presence clusters for the best model in the Model 3 series is shown in Fig. 2.4. Rather surprisingly, the clusters each seem restricted to a different state. Cluster 1 (red dots) is mostly in Andhra Pradesh, cluster 2 (blue) is mostly in Karnataka and cluster 3 (grey) is mostly in Tamil Nadu. Such a restriction of clusters (based on environmental variables) to particular states indicates that the environmental conditions of those states differ consistently one from the other. It is possible that such environmental conditions in turn determine rather different epidemiologies of BTV in the three states perhaps, for example, involving different key vector and/or host species.

This result also highlights the potential danger of extrapolating any model based on one state's data to other, even adjacent states that may (and in the present case clearly do) experience different environmental conditions.

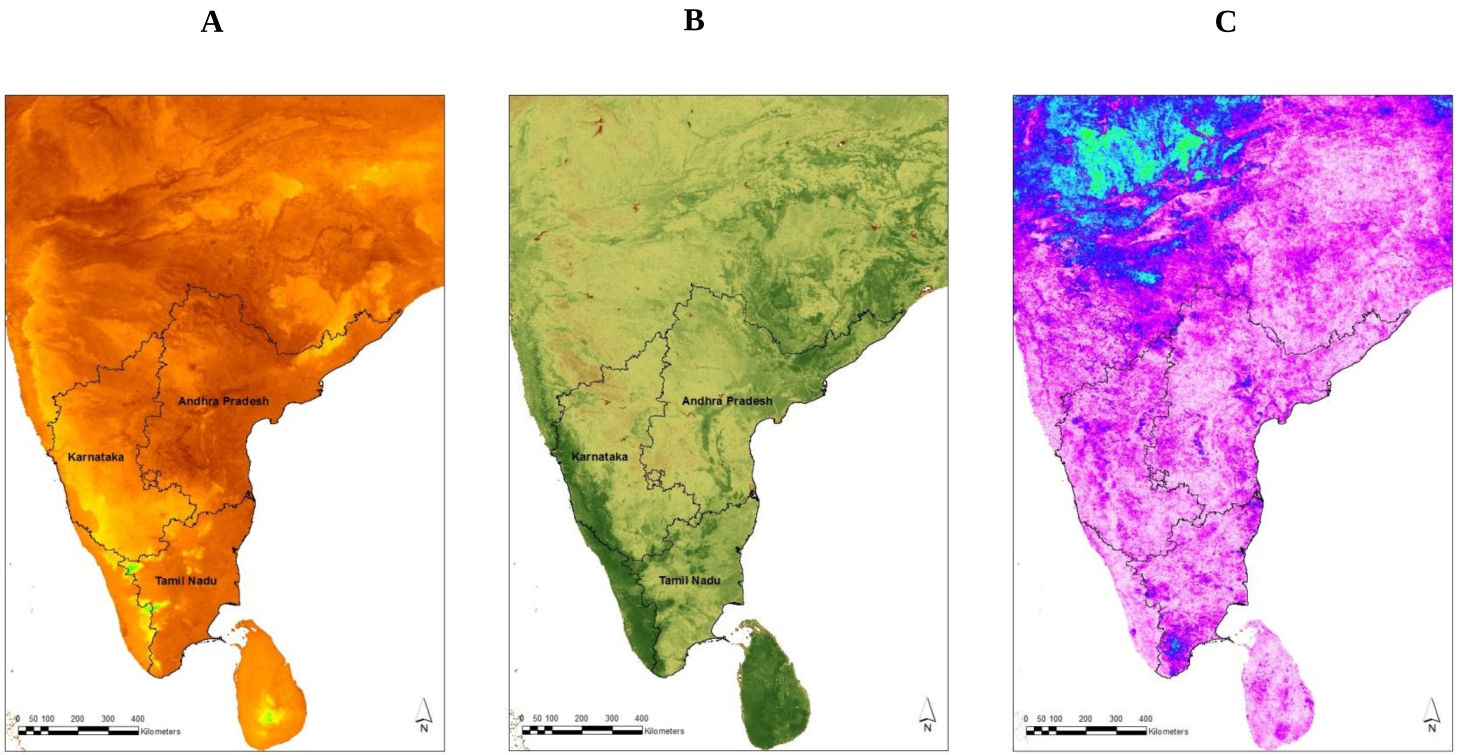


Figure 2.1: Maps of potential explanatory discriminating variables affecting presence and absence of bluetongue in South India: (A)Maximum day time LST; the maximum temperature is less along the western region (B) Mean NDVI; the NDVI is high (darker green) along the western ghat region and in a few areas in Tamil Nadu, Central and North-Eastern Andhra Pradesh and (C)Triannual amplitude day time LST; the tri-annual amplitude of temperature is low(white to light pink color) along western region and high in other regions(dark pink to blue). (Scharlemann et al., 2008).

	Kappa	Sensitivity	Specificity
Model 1			
Accuracy statistics	0.54± 0.048	0.79± 0.033	0.74 ± 0.037
Validation statistics	0.20 ± 0.039	0.69 ± 0.094	0.66 ± 0.027
Model 2			
Accuracy statistics	0.84 ± 0.036	0.93 ± 0.02	0.94 ± 0.014
Validation accuracy statistics	0.67 ± 0.063	0.87 ± 0.062	0.90 ± 0.011
Model 3			
Accuracy statistics	0.84 ± 0.026	**0.97 ± 0.013**	**0.96 ± 0.013**
Validation accuracy statistics	0.64 ± 0.035	0.88 ± 0.069	0.91 ± 0.012

Table 2.1: Accuracy and validation accuracy statistics (Kappa, sensitivity and specificity) for the three model series. See text for the conventional interpretation of the values of kappa. Model 2 and 3 have excellent performance on the training set compared to model 1. Model 1 performed poorly on the validation set compared to model 2 and 3. Overall Model 3 performed marginally better than the model 2.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Sample Size**
P1	7.27	1.5	303.05	3.26	0.24	1.26	4.87	0.39	21.90	308.09	62
P2	11.0	1.31	303.66	3.32	0.25	1.25	5.35	0.40	33.36	310.18	103
P3	7.2	1.03	301.08	3.21	0.30	1.23	5.52	0.47	26.97	308.42	35
A1	15.56	1.95	301.73	3.54	0.24	0.96	5.32	0.41	45.26	309.27	144
A2	7.5	1.22	298.82	3.34	0.34	1.65	5.66	0.54	20.25	304.61	32
A3	3	0.87	297.54	3.62	0.45	1.47	3.74	0.70	9.37	301.11	24
Mean P	**9.18**	**1.32**	**303.02**	**3.28**	**0.26**	**1.25**	**5.23**	**0.41**	**28.69**	**309.22**	**200**
Mean A	**12.76**	**1.71**	**300.76**	**3.52**	**0.28**	**1.13**	**5.18**	**0.46**	**36.95**	**307.54**	**200**

Table 2.2: Mean values of the top ten ranked variables from the best bluetongue model with lowest AIC$_C$ among the 100 bootstrap models. The mean values for the top ten variables for three presence (upper rows) and three absence clusters (lower rows). The last two columns show the mean values for Presence (P) and Absence (A) respectively. See Table 2.3 for the key to variable names.

Variable	**Expansion**
1	nLST variance, degrees K^2
2	Bi-annual amplitude of nLST, degrees K
3	Minimum dLST, degrees K
4	Bi-annual phase of nLST, decimal month (0 = January)
5	Mean EVI
6	Tri-annual phase of nLST, decimal month (0 = January)
7	Annual phase of nLST, decimal month (0 = January)
8	Mean NDVI (no units)
9	Variance of dLST, degrees K^2
10	Mean of dLST, degrees K

Table 2.3: Key to variable names for the best model among the 100 bootstrap models.

Model rank	Top ten Variables in Model 1	Top ten Variables in Model 2	Top ten Variables in Model 3
1	Maximum nLST	Minimum nLST	Phase of tri-annual cycle of nLST
2	Maximum dLST	Phase of triannual cycle nLST	Variance of nLST
3	amplitude of biannual dLST	Variance nLST	Amplitude of bi-annual cycle of nLST
4	Amplitude of biannual nLST	Minimum dLST	Minimum nLST
5	Mean of dLST	Phase of triannual cycle of dLST	Maximum dLST
6	Amplitude of tri-annual nLST	Mean MIR	Minimum dLST
7	Phase of biannual cycle of dLST	Phase of tri-annual cycle of MIR	Phase of tri-annual cycle of dLST
8	Phase of annual cycle of NDVI	Mean nLST	Variance of dLST
9	Phase of biannual cycle of nLST	Phase of biannual cycle of nLST	Mean of dLST
10	Minimum nLST	Maximum dLST	Phase of tri-annual cycle of NDVI

Table 2.4: Mean ranks of the top ten variables from all the 100 bootstrap models in the three Model series, 1, 2 and 3.

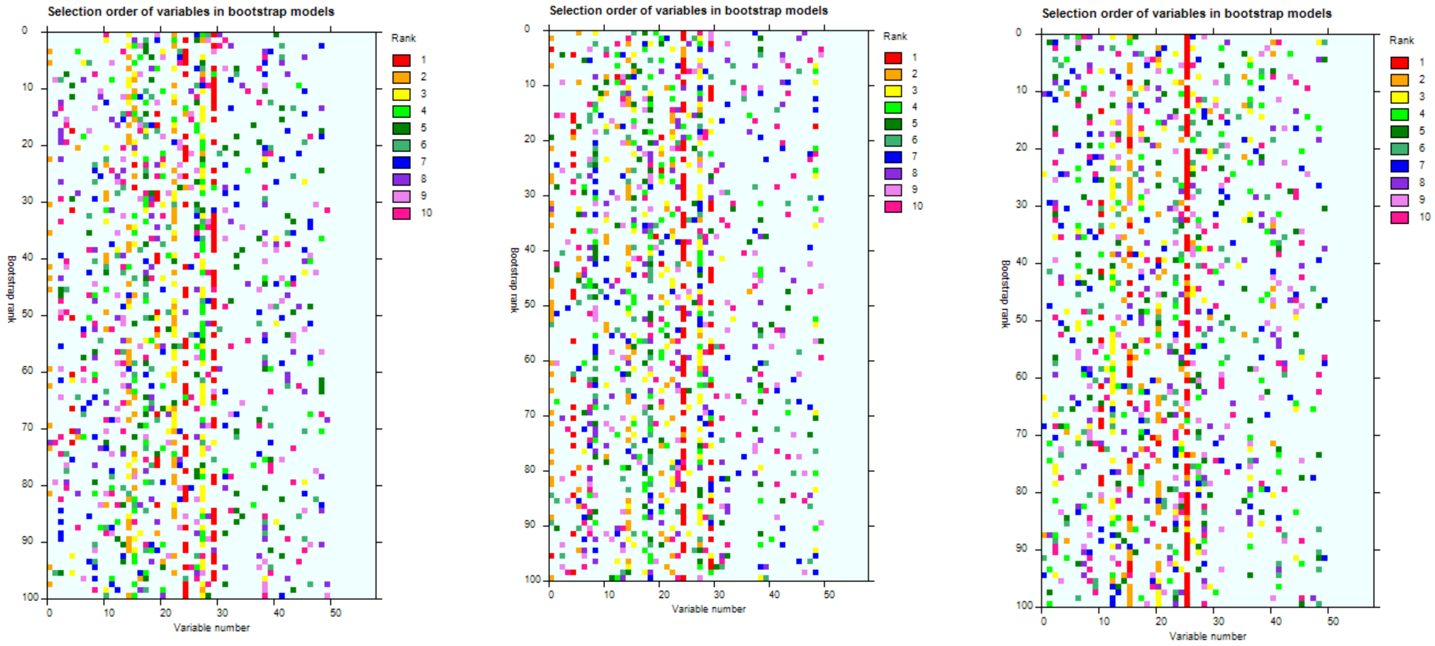


Figure 2.2: Graphical representation ('rainbow plots') of the ranks of predictor variables to show how often any particular variable was selected across the 100 bootstrap models for the Model 1, 2 and 3 series of models. Each row i these figures refers to a single bootstrap model (arranged in rank order, with the best model at the top), and each column to one of the predictor variables. In any single model the top predictor variable is colour coded red, the second most important variable is coloured orange and so on, on a rainbow colour scale (hence the description 'rainbow plot' for such images) – see legend in each plot for the colours. There is a predominant red line (phase of nLST tri-annual cycle) in Model 3, indicating not only that this variable was frequently selected, but also that it was often selected first in the various bootstrapped models. There are fewer signs of an overall dominant variable in the other two plots.

		Predicted category						
Observed category		P1	P2	P3	A1	A2	A3	Tot.
	P1	**52**	5	3	0	2	0	62
	P2	9	**87**	7	0	0	0	103
	P3	1	4	**29**	0	1	0	35
	A1	1	1	0	**139**	2	1	144
	A2	0	1	4	0	**27**	0	32
	A3	0	0	0	0	1	**23**	24
	Tot.	63	98	43	139	33	24	**400**

Table 2.5: Model accuracy matrix of the best model (pseudo absence with 3 presence and absence clusters each) among 100 bootstrap models. Model accuracy matrix of the best model (pseudo absence with 3 presence and absence clusters each) among 100 bootstrap models. The observed categories are in rows and predicted categories in the column. For a perfect model fit, all the numbers should be on the diagonal, with no off-diagonal entries. P1-P3 is the presence categories and A1-A3 is the absence categories.

Category	**%Correct**	**%Producer's Accuracy**	**%Consumer's Accuracy**
P1	83.87	83.87	82.54
P2	84.47	84.47	88.78
P3	82.86	82.86	67.44
A1	96.53	96.53	100
A2	84.38	84.38	81.82
A3	95.83	95.83	95.83

Table 2.6: Overall accuracy statistic for the best model among 100 bootstrap models. The overall kappa accuracy of this model is 0.86 and AUC: 0.9987.

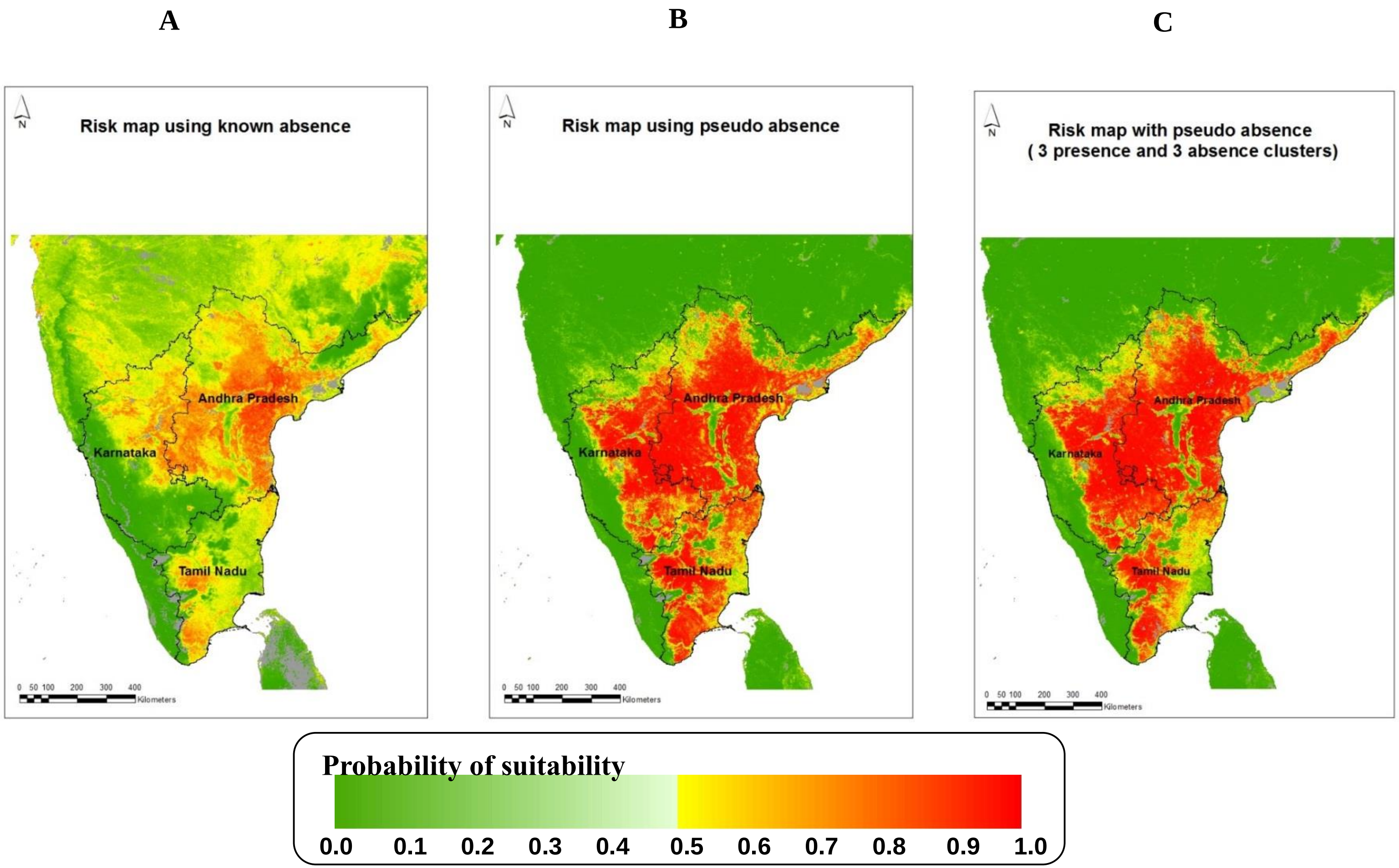

A
B
C
Risk map using known absence
Risk map using pseudo absence
Risk map with pseudo absence (3 presence and 3 absence clusters)
N
Andhra Pradesh
Karnataka
Tamil Nadu
0 50 100 200 300 400
Kilometers
Probability of suitability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Figure 2.3: Risk map for bluetongue in south India using Fourier processed MODIS variables with (A) Model 1 (known absence with 1 presence and 2 absence clusters), (B) Model 2 (pseudo absence with 1 presence and 2 absence clusters) and (C)Model 3 (pseudo absence with 3 presence and 3 absence clusters). The probability of suitability is on a scale from zero to one. Probabilities from 0.0 to 0.49 are coloured green (darker to lighter green) indicating predicted absence of disease. Probabilities from 0.50 to 1.0 are coloured yellow through dark red indicating conditions predicted suitable for bluetongue.

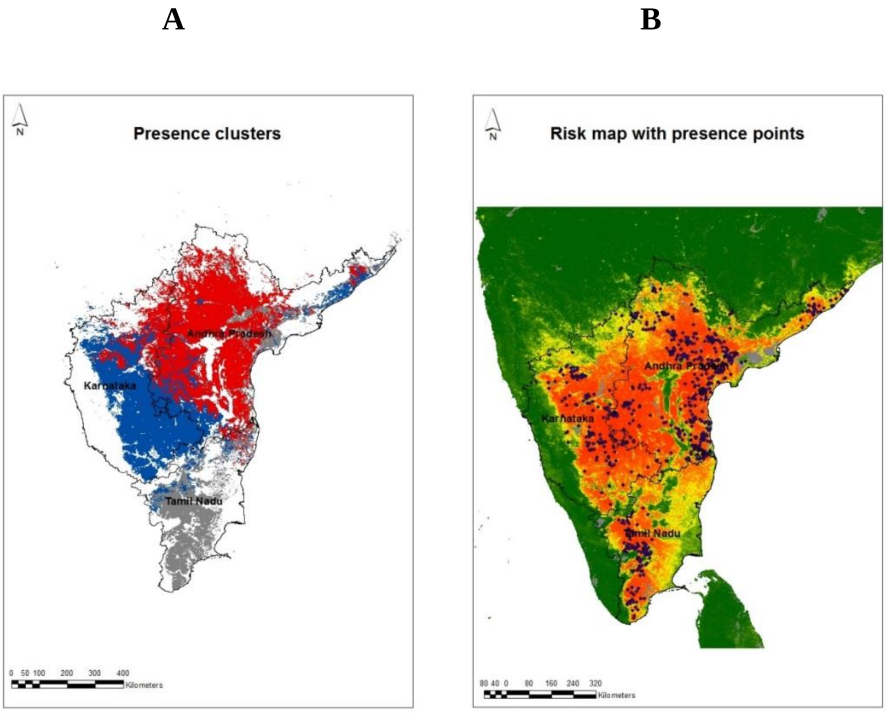


Figure 2.4: (A) Distribution of clusters in one of the Model 3 series of models, with 3 presence and 3 absence clusters. The absence clusters are shown in white. Notice that each cluster is relatively restricted geographically, one to each state. (B) Risk map developed with Model 3 series of models with bluetongue presence points (blue dots).

2.4 Discussion

Although there is high seroprevalence of bluetongue in India, with regular outbreaks occurring in South India, there are no studies to identify risk areas at the village level. Most studies concentrate on the molecular aspects of the virus or sero-prevalence. The district level forecast system predicts the most likely occurrence of bluetongue in districts and there is no risk map at village level. The risk map generated from this work using temporally Fourier processed remotely sensed variables discriminates between the presence and absence areas with high model accuracy for both training and test data.

Overall the model with pseudo absence points and with three presence and three absence clusters performed better than the other two models considered here. The model based on known absence points (villages never recording a BTV outbreak) gave a poor fit compared to models with pseudo absence points. The reason for this is not completely clear, but several suggestions were given in the Introduction to this Chapter. It is possible that the villages are suitable for BTV but simply, by chance, have not experienced BTV yet. Or they may have experienced unrecorded outbreaks of the disease. The minimum distance rule used to generate pseudo-absence data means that absence sites are environmentally more different from any presence site than are likely to have been the absence villages in the original dataset – some of which may have been very close to presence villages, both geographically and environmentally. Selecting pseudo-absence sites in this way may artificially inflate overall model accuracy (because absence sites are all very different from presence sites; hence discriminating the two sorts of sites will be easy), and this should also be considered when comparing models with different sorts of absence sites. The averaged bootstrap model accuracy (sensitivity =0.97, specificity=0.96) is very high and the validation statistics on "out of fit" test data (sensitivity =0.88 and specificity=0.91) are also very good considering the very different environmental conditions across the three states under study. Most of the models for predicting bluetongue or *C.imicola* presence and absence in other countries have been validated internally and the percentage of correct classification ranged from 75% to

95% (Calistri et al., 2003; Conte et al., 2003; Tatem et al., 2003; Wittmann et al., 2001).

In all three model series, temperature variables were predominant in the highest ranked variables (Table 2.4). The selection of temperature variables in discriminating presence and absence areas is as expected because temperature not only influences the different life stages of *Culicoides,* but also influences the extrinsic incubation period. The higher values of nLST variance in the P2 cluster compared to A2 and A3 (Table 2.2) shows the influence of temperature between the presence and absence groups, and therefore comparing overall means of a single variable of the presence and absence groups can be misleading. Instead it is the unique combination of variables (not the value of any single one) in each group that tends to determine disease presence and absence. The presence of different environmental conditions may be a contributing factor for presence of different epidemiological systems along with diversity of hosts and breeds. The detection of different groups is also supported by the fact that there are different *Culicoides* species identified in each state and different serotypes detected in the past (Sreenivasulu et al., 2004). However, studies on understanding the role of *Culicoides* species in transmitting different serotype/strains is lacking. The presence of three distinct presence and absence groups (Fig. 2.4 a) which corresponds to three states under study is supported by the fact that major areas of these states fall under different agro-ecological zones of the country. The present study also stresses the usefulness of Fourier variables in discriminating between the different areas of endemism of the disease.

NDVI variables were important in describing the distribution of *C. imicola* in temperate regions (Baylis et al., 1998; Baylis et al., 1999; Baylis et al., 2001; Tatem et al., 2003) whilst (Purse, et al., 2004) found that different variables were selected for each vector species modelled (NDVI for *C.pulicaris* & *C.imicola* and temperature variables for *C.obsoletus* group and *C.newsteadi*). However, NDVI along with minimum LST explained more variance (67%) compared to a model with just minimum LST (explaining only 40% of the variance) in a *C.imicola* abundance model (Baylis, et al., 1999). In another study the mean of Middle Infra-Red reflectance (MIR) was the most important variable in determining the presence of *C.imicola*, but NDVI was selected in an abundance model (Tatem, et al., 2003). In the present study, NDVI is higher (Fig. 2.1B) in areas where bluetongue has never been reported (Fig. 2.4b). The best risk map developed here shows that the areas along the Western ghat (Fig. 2.3C) are less suitable for bluetongue transmission. This region is covered by forest, has fewer sheep, and also fewer movements of livestock from other regions of South India than do other parts of the state. The Western ghat regions have not reported bluetongue outbreaks in the past twenty years or so (district level NIVEDI database on livestock diseases and Fig. 2.4B). The Eastern ghat region is similar in many respects to the Western ghat region and is also predicted by the model to be at very low risk of BTV infections. There is a common belief among researchers and field Veterinarians that bluetongue is endemic in the whole of South India, but both the data and the risk maps show that there are areas within each state that are very unlikely to experience BTV outbreaks because their environmental conditions are so unlike any of the sites from which BTV has been recorded to date. Although there is sero-prevalence of bluetongue in Kerala (Ravishankar et al., 2005), which forms part of Western Ghat, no clinical outbreaks of the disease have been reported in the past 20 years. One of the reasons for the absence of clinical disease is due to low sheep population and high goat population (goats are relatively resistant to BTV compared to sheep or due to host specificity of potential vectors) in Kerala.

The map of the presence clusters shows that a considerable area is covered by each cluster which is, nevertheless, relatively distinct geographically from the other clusters. This suggests that disease transmission may be different in the different areas, perhaps involving different vectors or vector complexes, different hosts and possibly different environments for transmission, or any combination of these effects. Thus there is an urgent need for systematic vector and host competence studies in the region. In conclusion, discriminant analysis throws an interesting light on potentially different epidemiological situations, and raises relatively precise questions for future studies to address.

Chapter 3

Role of intrinsic and extrinsic factors in driving temporal patterns in bluetongue outbreaks in India: a Bayesian time series regression approach

3.1 Introduction

In Africa and India, bluetongue occurs every year with varying severity (Coetzee et al., 2012; Prasad et al., 1992). Multiple serotypes circulating in India (Sreenivasulu et al., 2003) and Africa (Coetzee et al., 2012) are thought to be transmitted by many different species of *Culicoides* (Ilango, 2006). Limited knowledge of the vectorial capacity of indigenous species limits our understanding of the epidemiology of disease in India. Elsewhere, annual variability of bluetongue outbreaks has been linked to the effects of previous climate events (Purse, et al., 2004) while African Horse sickness outbreaks have been linked to El-Nino (Baylis et al., 1999b).

In India, Andhra Pradesh is the state most severely affected by Bluetongue virus and its surveillance system is better than in the other states (Ahuja et al., 2008). Bluetongue in Andhra Pradesh varies in severity across months and years.

A forecasting model that predicts the presence and absence of bluetongue outbreaks at district level from environmental covariates (livestock demography, climate and land use pattern) exists for India (www.nadres.res.in). The forecasting ability of these environmental models has not been systematically evaluated on "out of fit" data for bluetongue in India accounting for the past dependency of outbreaks (temporal autocorrelation) and using methods to handle count data (Poisson regression).

Bluetongue varies on a seasonal and annual basis (Fig. 3.1A) with more outbreaks in certain months and years than others. Seasonal and annual variation in bluetongue outbreaks can be due to interaction between a range of extrinsic and intrinsic factors, discussed in the following sections.

3.1.1 Extrinsic factors

The maintenance of BTV is either due to the presence of adult *Culicoides* (potential vector) throughout the year or the presence of virus in the blood of infected or reservoir hosts. The possibility of the latter is doubtful because detectable viraemia in cattle or other reservoir hosts is restricted to a maximum of about 50 days following patent infection (Bonneau et al., 2002). There are reports of the presence of adult *C.imicola* throughout the year in North Karnataka (Bhoyar et al., 2012), and such continuous presence requires suitable climatic conditions (Mellor et al., 2000). Temperature influences the development and survival rates of *Culicoides,* and viral replication within the adult vectors, all of which in turn govern the transmission of BTV (Mellor et al., 2000). Additionally, rainfall can also influence larval development, survival and the abundance of *Culicoides*. The influence of flooding or drought on the immature stages depends on the *Culicoides* species. The pupae of *C.imicola* drown when breeding sites are flooded (Nevill, 1967), whereas pupae of the *Pulicaris* group in South Africa are tolerant of waterlogged breeding sites and even prefer them, because the pupae can float on the water surface (Nevill et al., 2007).

Extreme weather events caused by El-Niño may be responsible for inter-annual variability of rainfall in India. An El-Niño effect on outbreaks of African Horse sickness in Southern Africa has been detected (Baylis et al., 1999b).

3.1.2 Intrinsic factors

A disease may be absent from an area even when extrinsic conditions are suitable if some other factor in the transmission cycle comes into play. Herd immunity is one of several intrinsic factors responsible for the waxing and waning of infection rates even in the absence of climatic seasonality. In the case of bluetongue, several serotypes may be co-circulating in the same place (Coetzee et al., 2012) and not all serotypes cause severe disease. Immunity is serotype specific (Schwartz-Cornil et al., 2008) and any cross protection (which in the case of BTV occurs only with serotypes with similar Virus Protein 2, nucleotide sequences (Maan et al., 2007)) may reduce BTV outbreaks periodically, until such immunity diminishes or is lost through natural recovery, or death. Therefore the occurrence of pathogenic serotypes (which cause more severe disease) may be determined by the waxing and waning of more general (Virus Protein 2 determined) herd immunity.

New births of course replenish the stock of susceptible animals and, in general, the natural periodicity of immune-driven disease outbreaks is a function of the host population’s birth rate (Anderson et al., 1992).

Herd immunity is potentially important epidemiologically only when infection rates exceed some threshold level, such that significant proportions of local herds are immune-protected when BTV transmission rates increase seasonally. The precise level of this threshold varies from one disease to another and its impact on disease outbreaks will depend upon the degree of seasonality in each transmission site. Unfortunately, few surveys have been carried out in India to date to investigate local levels of immunity to BTV so that this and several other intrinsic factors may be playing an as yet un-quantified role in BTV transmission.

3.1.3 Forecasting and Early warning system

Forecasting and early warning systems have been used for a number of vector-borne diseases (Hii, et al., 2012), but there is no such system for bluetongue in India, except for a district level presence and absence forecasting system (www.nadres.res.in). The development of a forecasting system for any vector borne disease relies on the quantification of past outbreaks with vector and/or climatic variables using techniques accounting for temporal autocorrelation. Any model developed for forecasting vector-borne diseases should be able to capture both short term and long term changes in the dependent variable and should be robust enough to make projections into the future on the basis of predictor variables from the present or recent past.

Understanding the role of intrinsic and extrinsic factors that determine the severity of BT outbreaks in AP requires analytical methods that can deal with both the temporal dependence and non-normality of the outbreak data. In this chapter the role of weather and long term climate variation associated with El-Niño is investigated to seek answers to the following questions:

1. What are the relative roles of prior monthly conditions of temperature and rainfall, including monsoon conditions, in determining seasonal patterns in bluetongue outbreaks?
2. Are BTV outbreaks in India cyclical, with periodicities of greater than one year?
3. Once obvious environmental impacts are accounted for, is there any residual variance or periodicity in disease outbreaks that may indicate an important role of intrinsic factors such as herd immunity?
4. Can future BT outbreaks be forecast adequately using a model parameterised on past outbreak data?

3.2 Materials and methods

3.2.1 Disease data

Bluetongue outbreak data: District level (admin-2) monthly BT outbreak data (1992-2009) were provided by NIVEDI (National Institute of Veterinary Epidemiology and Disease Informatics), formerly known as PD_ADMAS (Project Directorate on Animal Disease Monitoring and Surveillance), which maintains the livestock diseases database for India and collates outbreak data every month from different sources. The analysis was restricted to data from Andhra Pradesh because this state has regular reporting and better surveillance (Ahuja et al., 2008) compared to other states of India. The district level data were aggregated across Andhra Pradesh to calculate the sum of the bluetongue outbreaks every month as the state-wide dependent variable (Fig. 3.1). The time series was divided into a training set (1992-2004) and a test set (2005-2007).

3.2.2 Weather and El-Niño data

Temperature and precipitation data

Concurrent monthly mean, minimum and maximum temperature and precipitation estimates for Andhra Pradesh were extracted from the CRU TS3.10 dataset from the Climatic Research Unit (CRU) (http://www.cru.uea.ac.uk/data), University of East Anglia (Harris et al., 2014). All the temperature and precipitation variables were mean-centred (by subtracting the synoptic monthly means calculated for the period 1992 to 2009 from the raw data) to give twelve monthly figures for each year. Centring of predictor variables helps to improve the efficiency of MCMC sampling (McCarthy, 2007; Searle et al., 2013) and the intercept is therefore the expected value of the response variable when all the predictor variable values are set to their means. The intercept when the independent variables are not mean-centred is the expected values of the response variable when all the predictor variables are set to zero.

El-Nino -3 (sea surface temperature)

The coastal warming of the Pacific Ocean that is linked to anomalies in global climate is known as “El-Nino” (Trenberth, 1997). The fluctuation of the atmospheric pressure over the ocean is known as the “Southern oscillation”. The Pacific ocean warming and the fluctuations in the atmosphere together are known as ENSO (El Nino-Southern Oscillation). The El-Nino corresponds to a warm phase of ENSO and La Nina corresponds to a cold phase of ENSO, although commonly (and incorrectly) both the cold and warm phase of ENSO are referred to as El-Nino. The Nino 3 region covers latitudes 50^0N to 5^0S and longitudes 90^0W-150^0W. The monthly sea-surface temperature (SST) data (1992-2009) were obtained from the Japan Meteorological Agency (JMA) (Ishii et al., 2005). The JMA analyses the SST data to calculate SST anomalies. Five-month running means of monthly SST anomalies are used to identify periods which have anomalies of >= ±0.5^0C and these periods define the JMA Nino 3 time series; El-Nino (positive deviation) and La Nina (negative deviation). A lower threshold of ±0.4^oC is used for the JMA Nino 3.4 time series (Trenberth1997).

3.2.3 Statistical methods

To determine whether there were significant autocorrelations (BTV outbreaks at a given time depending on prior outbreaks) and to identify whether the series is stationary (i.e. has a constant mean and constant variance), autocorrelation functions (ACF) and partial autocorrelation functions (PACF) were examined.

The ACF is the cross-correlation of the time series with itself as a function of time lag, k, between time points and lies between -1 to +1. The series is stationary if the ACF falls from one to zero immediately and non-stationary if the ACF falls from one to zero gradually. The statistical significance of the ACF at different lags can be tested using the Q-test or the Ljung-Box statistic (Chatfield, 2013).

The PACF measures correlation between BT outbreaks that are k time periods apart, after controlling for all correlations at intermediate lags. In other words partial correlation is a conditional correlation, that is, a correlation between two variables after their mutual linear dependency on the intervening variables Y_{t-1}, $Y_{t-2}Y_{t-k+1}$ has been removed. PACF is used to identify the order of autoregressive model in ARIMA (Autoregressive Integrated Moving Average) models or other models with non-Gaussian data. PACF was used in this analysis to identify the order of the autoregressive term.

Cross-correlation functions (CCF) were examined to identify the lagged relationship and dependency between the environmental time series and the BTV outbreaks. Performing cross correlation analysis on a raw time series is not advised (Chatfield, 2013) when the two time series are serially correlated and the correlation co-efficient on raw series can be misleading. Pre-whitening is performed when the driving variable (meteorological variables in this case) is serially correlated and cross-correlation using the raw series will not give the exact correlation between the driving and outcome (BTV outbreaks) variable. An ARIMA model is fitted to the driving variable and the same model is used to ‘filter’ the outcome time series. The residuals of the driving variable model and the filtered time series are then used to calculate the correlation co-efficient at different lags. The lags at which the relationship between temperature and rainfall and BTV outbreaks were significant were used in the model building process (Chatfield, 2013). The maximum lag at which the CCF is calculated is restricted to *$10*\log_{10}$ (N/m)*, where N is the number of observations in the time series and m is the number of series (Ripley 2002). For example in this case we have 216 observations and 2 time series, so the maximum lag at which the CCF is calculated is $10*\log_{10}$ (216/2) = 20.33, so the CCF will be plotted for ~ ±20 lags. The significance for ±20 lags in the interval is determined from ±2/√N= ±0.14 (corresponding to 2 S.E.). The lags at which significant correlations were identified were offered to the model building process.

Periodicities in data can also be identified using time series techniques in the frequency domain - also known as spectral analysis. Fourier transformation is a method used to identify periodicity in the data (Chatfield, 2013), but assumes that the mean, variance, and temporal dependence between data points all remain constant over time (referred to as stationarity). However, the condition of stationarity is usually violated in disease time series and often in climate time series, which are then called non-stationary (Chaves & Pascual, 2006). Vector borne disease time series data are often non-stationary due to seasonality in potential vectors transmitting the disease and also due to waxing and waning of herd immunity. The time series under study can be made stationary by differencing, detrending or other transformation techniques (Chaves & Pascual, 2006) before Fourier transformation. The disadvantages of Fourier transformation can be overcome by employing non-stationary time series tools such as wavelet analysis (Cazelles et al., 2008). Wavelet transformation is performed either by a discrete method (Discrete Wavelet Transform, DWT) or a Continuous (Continuous Wavelet Transform, CWT) method (Grinsted et al., 2004). Both convolve a time series with a localised wavelet function and extract from the product information about the local (in time) power of the signal at the characteristic wavelength of the wavelet function. By effectively varying this wavelength the localised power in the signal at a variety of wavelengths (= periodicities) can be calculated. The only difference between DWT and CWT is that the former is applied at discrete intervals of time and wavelengths (hence the output wavelet diagram has a 'blocky' appearance) and the latter is applied continuously (and hence the output diagram has a smoother looking appearance). The DWT is computationally simpler and is appropriate for data sampled at relatively long intervals; the CWT is more appropriate for more frequently sampled data and when dominant periodicities change gradually over time. The CWT was used in the present analysis.

Wavelet analysis has been used in epidemiological time series (Cazelles et al., 2007) to identify the periodicities in the data which can be included in the model using harmonics or other smoothing techniques. Wavelets can also be applied to two time series simultaneously, to identify the dominant periodicities in the time series and their cross correlation using cross-wavelets (wavelet coherence). This technique will be helpful in the case of bluetongue where there is not only within year seasonality but also inter-annual variability in outbreaks. Cross wavelet analysis usually produces two graphs, one the 'cross wavelet transform' that shows the common power in the two signals (effectively the product of the local amplitudes) and the other the 'cross wavelet coherence' which is effectively the localised correlation coefficient (squared, hence on the scale of zero to 1.0) between the two signals. To either graph (more commonly the former) may be added arrows indicating the localised phase lag between the two signals. Consistent phase lags within regions of high joint power (i.e. all localised arrows pointing in the same direction) are a strong indication that one process (e.g. rainfall) is driving the other (e.g. disease outbreaks) at the lags corresponding to the local phase difference (N.B. turning phase differences into absolute lags for modelling purposes is scale dependent).

The wavelet transform is calculated from:

$$W_x(a,\tau) = \int_{-\infty}^{+\infty} x(t)\phi *_{a,\tau}(t)dt$$

Where * denotes the complex conjugate

x (t) represent the time series of interest.

W_x (a,τ): wavelet co-efficients represent the contribution of scales or widths (the a values) to the signal at different time positions (the τ values)

ϕ(t): is known as "mother wavelet", the scale of which (a) is changed (discretely or continuously for the DWT or CWT respectively).

Time series analysis

Linear regression methods assuming residuals of the analysed time series are uncorrelated (Selvaraju et al., 2013) are inappropriate in the case of infectious diseases, when current outbreaks are dependent on the past outbreaks. Nevertheless, there are numerous examples in other VBD's which incorporate temporal dependency in the time series with Gaussian outcomes (Gharbi et al., 2011; Luz et al., 2008; Wangdi et al., 2010) using the popular Box-Jenkins method (Helfenstein, 1986). Purely autoregressive time series models have been extended to include weather variables (Gharbi et al., 2011; Helfenstein, 1991) and their lags, and also to account for seasonality (Zhang et al., 2010). However, these autoregressive time series methods are well established for Gaussian outcomes, whereas models for non-Gaussian count data are less well developed in environmental epidemiology (Bhaskaran et al., 2013) and very few in infectious disease epidemiology (Chou et al., 2010; Fernández et al., 2009; Lu et al., 2009) as discussed in the Introduction chapter.

The assumption of the Poisson distribution (the equivalence of variance and mean) is often violated by epidemiological time series data which tend to be over dispersed (variance > mean) (Heinen, 2003). This over dispersion may be due to autocorrelation in the dependent variable in Poisson regression or may be due to changes in the epidemiological system itself (seasonality, herd population structure, movement and immunity, variation in serotypes etc.) (Altizer et al., 2006). Testing for residual autocorrelation and the presence of residual over dispersion is important to avoid errors in estimates (Scrucca et al., 2014). Residual autocorrelation can be accounted for by using additional autoregressive terms.

Time series techniques should account for both short-term (seasonal) and long-term (inter-annual) dependencies. Short term dependency in the response variable may be due to extrinsic factors discussed above and can be accounted for by inclusion of covariates and by autoregressive terms. Long term dependency may be due to inter-annual climate variability or the presence of herd immunity which can result in bias in estimates if not accounted for in a Poisson model.

Relationships between the monthly numbers of BT outbreaks and environmental predictors were quantified using a seasonal Generalised Linear Mixed Model with Poisson errors, implemented in a Bayesian framework (Sanders et al., 2011). In the event of finding significant autoregressive structure (AR (1)) in the PACF and ACF plots, a Bayesian Poisson model with autoregressive errors was fitted.

The number of BT outbreaks (y_t) observed in month *t* with corresponding meteorological variables x_{nt} was assumed to follow a Poisson distribution

The probability function for Y is given by Eq. (1).

$$\Pr(Y = y / \mu) = Poisson(\mu_t)$$

$$\log(\mu_t) = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_n x_{nt} + \varepsilon_t \qquad (1)$$

$$\varepsilon_t = \sum_{i=1}^{p} \rho_i \varepsilon_{t-1} + Z_t$$

$$Z_t \sim N(\mu\varepsilon, \tau\varepsilon)$$

priors for different parameters

$$\beta \sim N(\mu_\beta, \Sigma_\beta) \text{ - regression co-efficients}$$

$$\rho \sim dunif(-1,1) - \text{autocorrelation parameter}$$

$$\mu\varepsilon \sim N(0,0.0001) - \text{mean of autocorrelation}$$

$$\tau\varepsilon \sim gamma(1,1) - \text{precision of autocorrelation}$$

Model building process

An all subset approach was difficult to perform given the computational intensity of MCMC estimation. Therefore, all possible combinations of monthly lags identified in the cross correlation of rainfall (0, 1& 2 lags) and temperature (2, 4 & 8 lags) were first fitted individually to identify the best model based on reduced DIC (Spiegelhalter et al., 2002). Then, all possible combinations from the best model of rainfall lags and of temperature lags were fitted to identify the best monthly model, again based on reduced DIC. The residuals were tested for presence of any over dispersion using an overdispersion test (Scrucca et al., 2014) and for autocorrelation using the PACF.

Root mean square errors of the best models at each stage (monthly rainfall, monthly temperature, best combined monthly model) were calculated to evaluate the predictive power of the model. The final model was used to make a forecast of the "out of fit" data.

3.3 Results

The maximum number of bluetongue outbreaks (470) was observed in September 2005 in Andhra Pradesh. Monthly variation of bluetongue outbreaks, rainfall, maximum temperature and sea surface temperature (SST) are shown in Fig. 3.1. No trend is observed in bluetongue outbreaks over the study period (Fig. 3.1) (measurements do not increase or decrease systematically), so there is no need to difference the series for further analysis (a commonly employed technique for data that trend over time).

Monthly and yearly box plots (Fig 3.2 A) of bluetongue outbreaks shows seasonality and inter-annual variability respectively. Outbreaks usually start in August, peak in September and decline from October onwards. The yearly box plots show significant inter-annual variability.

Significant seasonality in rainfall is also observed due to the influence of both the South-West monsoon (June-September) and the North-East monsoon (October-December), and it also varies substantially between years.

Similarly, monthly maximum temperatures (Fig. 3.2C) start to increase from the month of February and stay high (over 40^0C) until the end of May (summer season) and then decline with the onset of monsoons from June onwards (Fig 3.2B).

The PACF plot (Fig 3.3B) of the BT series shows significant autocorrelation only at lag 1, indicated by the areas above the blue dotted horizontal line (95% confidence interval). In contrast there is strong significant autocorrelation in the rainfall and maximum temperature data (Fig. 3.4).

Cross-correlations between the pre-whitened environmental time series and filtered bluetongue shows that there were significant correlations at lag 0, 1 and 2 for rainfall (all positively correlated), lag 2 and 8 for minimum temperature (negatively correlated) and lag 2 and 8 for mean temperature (negatively correlated) (Fig. 3.5). Maximum temperature at lag 2 and 8 is negatively correlated with bluetongue outbreaks and positively correlated with maximum temperature at lag 4.

Cross-wavelet transformation of the BT and rainfall show matching of power at the dominant frequencies corresponding to a 12 month periodicity through much of the time series, whereas wavelet coherence (correlation) between BT and rainfall shows less consistent correlation at a periodicity of 12 months but a clear correlation in signals at 24-30 months periodicities (Fig. 3.6). Thus this area of the time/frequency plots has relatively low joint power (Fig. 3.6C) but a high correlation (Fig. 3.6D) between the rainfall and BT time series.

Cross wavelet and coherence analysis of the bluetongue outbreaks and El Niño 3 data reveals few areas of significant joint power (Fig. 3.7C) or coherence (Fig. 3.7D). This is unsurprising given the relatively large amounts of power at many different periods of < 1 year in both sequences, and the low amounts at periods of >= 1 year, surprisingly more noticeable in the El Nino series (Fig. 3.7A). There is strong correlation between maximum temperature and El Niño3 at 6 months' periodicity, and this is relatively consistent throughout the time series. The coherence of El Nino3 and rainfall is similar, but less pronounced (Fig. 3.8B).

Comparison of different BT models using only rainfall predictor variables identified the one using rainfall at lag 2 and the autoregressive error term (AR(1)) as the best, based on reduced DIC (Table 3.1). For those models using only temperature predictors, the best one used all the available lags (2, 4 and 8) (Table 3.1). Amongst all of the above models, however, the single overall best one was the simplest (rainfall at lag 2 plus AR (1) errors) and this model was subsequently used for making predictions on 'out-of-fit' data.

The mean co-efficient values and their credible intervals for best model with rainfall at lag 2 (Table 3.2) shows that monthly bluetongue outbreaks are significantly and positively associated with rainfall at lag 2. The fit of this model to the data is shown in Fig. 3.9. Comparison of models with and without temporal autocorrelation and covariates model shows that the model with covariate (rainfall at lag 2) and AR (1) outperforms the models with covariate only or AR (1) only (Table 3.3 and Fig 3.10), the latter of these two showing a reduction of 6000 DIC units compared with the former. The variance explained by the AR (1) only model is slightly more than the combined model (covariate and temporal autocorrelation), but far better than the covariate only model.

The best model (rainfall lag 2 and AR (1)) predicts a maximum of 300 outbreaks per month in 2005 (400 outbreaks were recorded), although the precise timing of this seasonal peak was not captured. The two following years (2006 and 2007) are correctly predicted to have fewer outbreaks than in 2005 but again there are quantitative (number of outbreaks) and qualitative (seasonality) differences between predictions and reality (Fig. 3.11). The value of the forecasting model is therefore questionable beyond predicting the overall severity of each year’s outbreak.

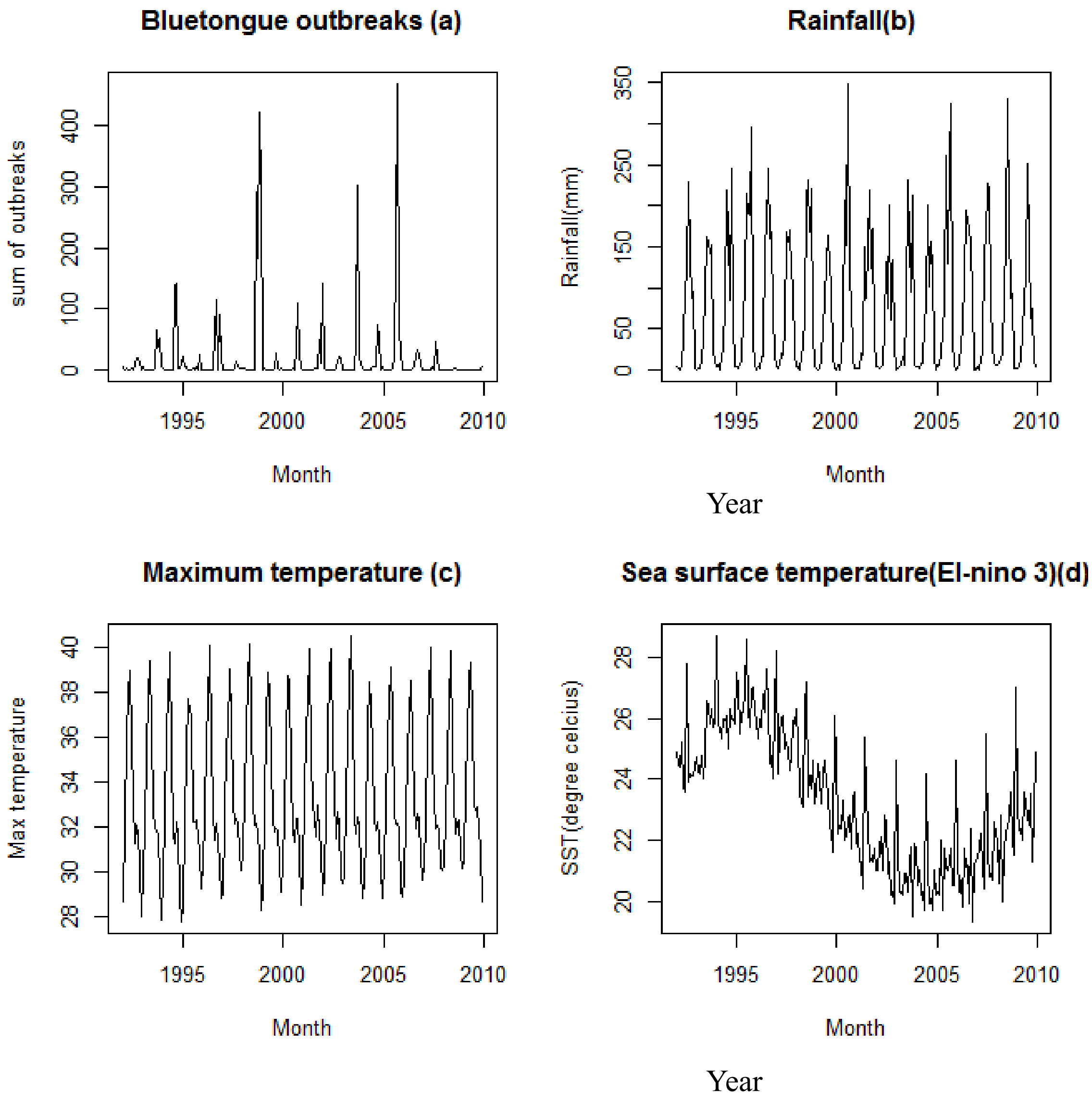


Figure 3.1: Time series plots of (a) monthly Bluetongue outbreaks (b) Monthly rainfall (mm) (c) Maximum temperature (^{0}C) and (d) sea surface temperature data(^{0}C) from 1992-2009 (El-Niño 3). The figures a-c shows the values for Andhra Pradesh and figure d shows the values of SST for the geographical area between 5^0 N-5^{0}S, 90^0-150^{0}W.

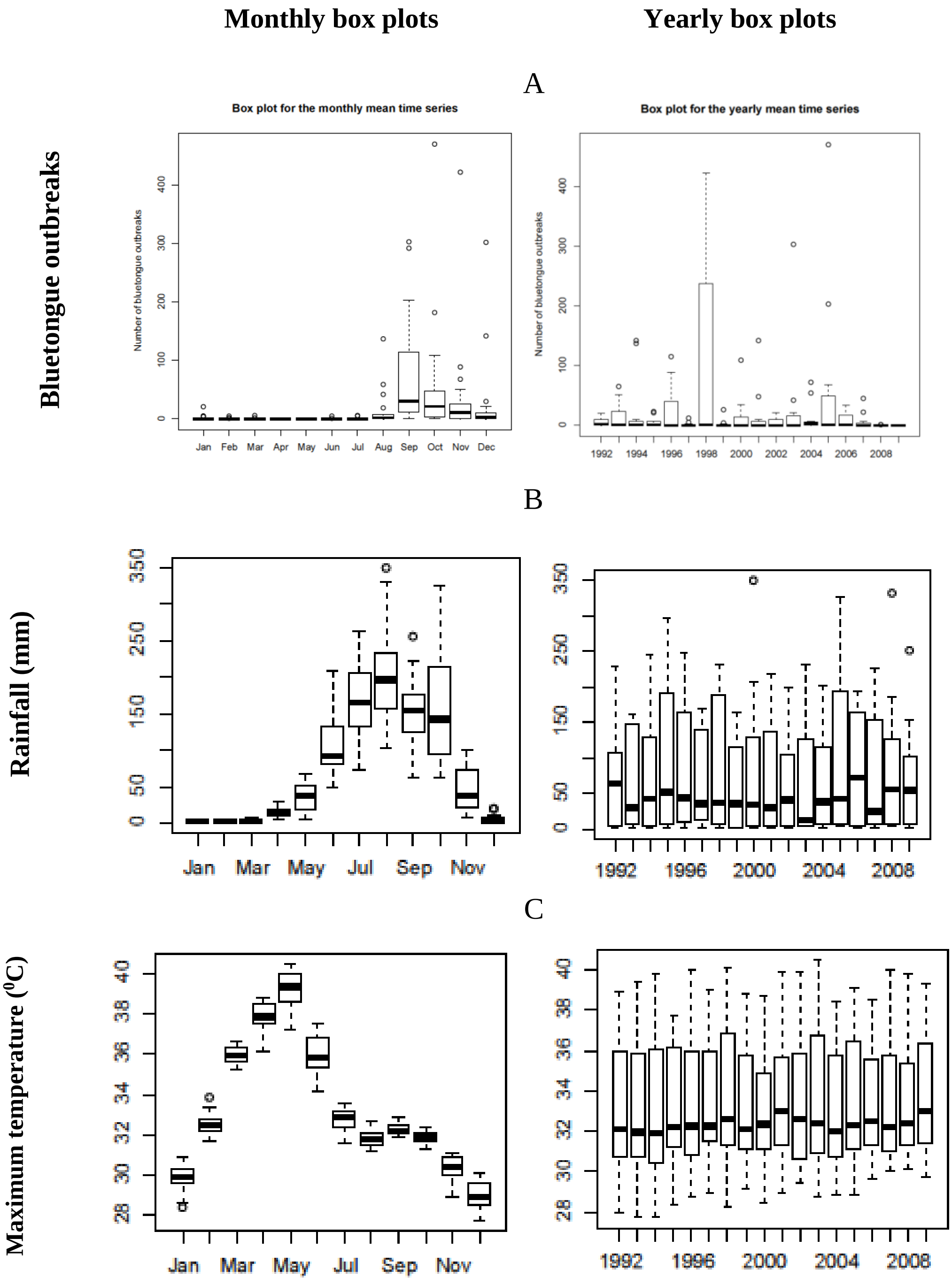
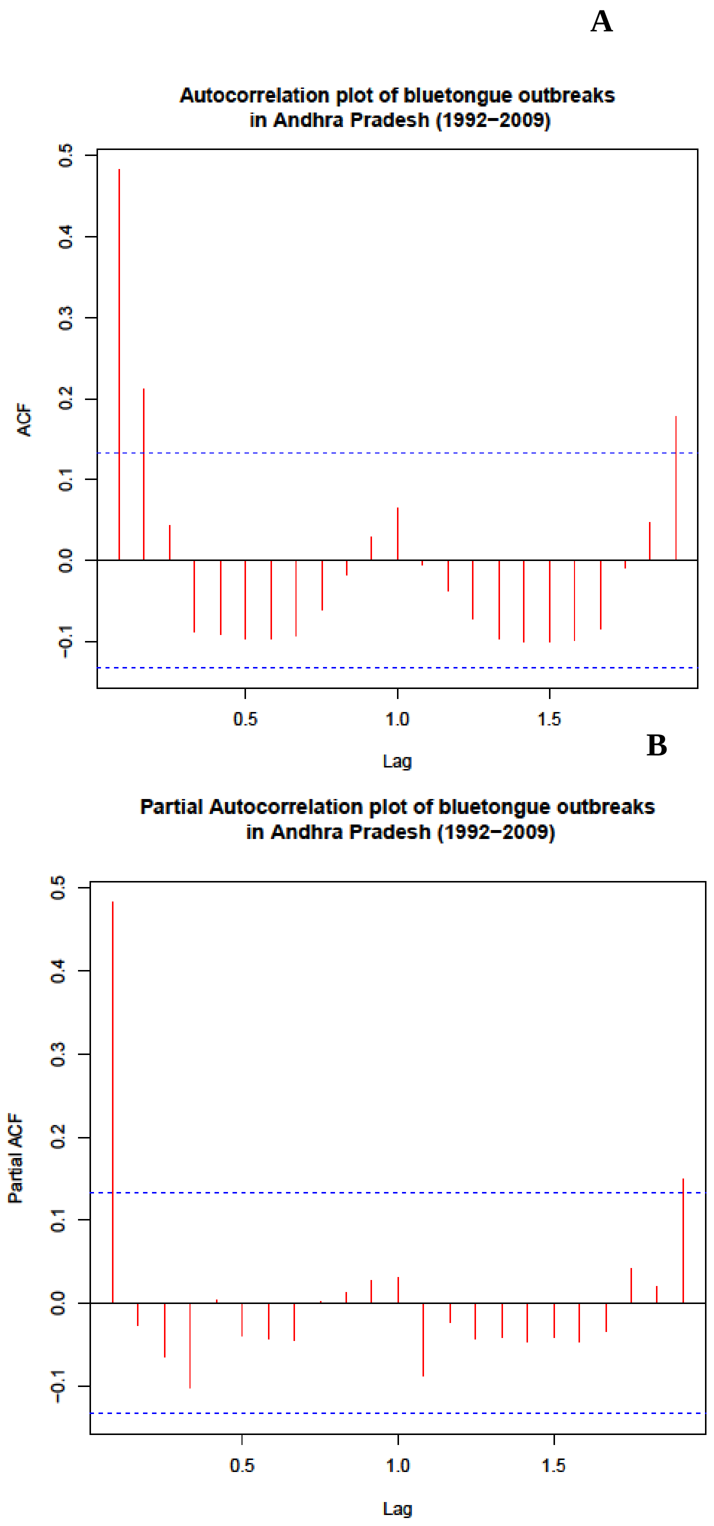


Figure 3.2: Box plots of the raw monthly and yearly mean bluetongue outbreaks and environmental variables. (A) Mean monthly and yearly bluetongue outbreaks (B) Mean monthly and yearly Rainfall (C) Mean monthly and yearly Maximum temperature. The box represents 50% of the values of the data and the whisker represents the minimum and maximum values in the data. The median is represented by horizontal line within the box. The box plots are used to display median, dispersion and skewness in the data (indicated by a median line not centred in the box).



Autocorrelation plot of bluetongue outbreaks in Andhra Pradesh (1992−2009)

Partial Autocorrelation plot of bluetongue outbreaks in Andhra Pradesh (1992−2009)

Figure 3.3: Plots of (A) Autocorrelation function (B) Partial autocorrelation function of bluetongue outbreaks. The X-axis gives the number of lags in years and, the y-axis gives the value of the correlation between -1 and 1. Blue dashed lines indicate the 95% confidence intervals, within which the correlation is non-significant.

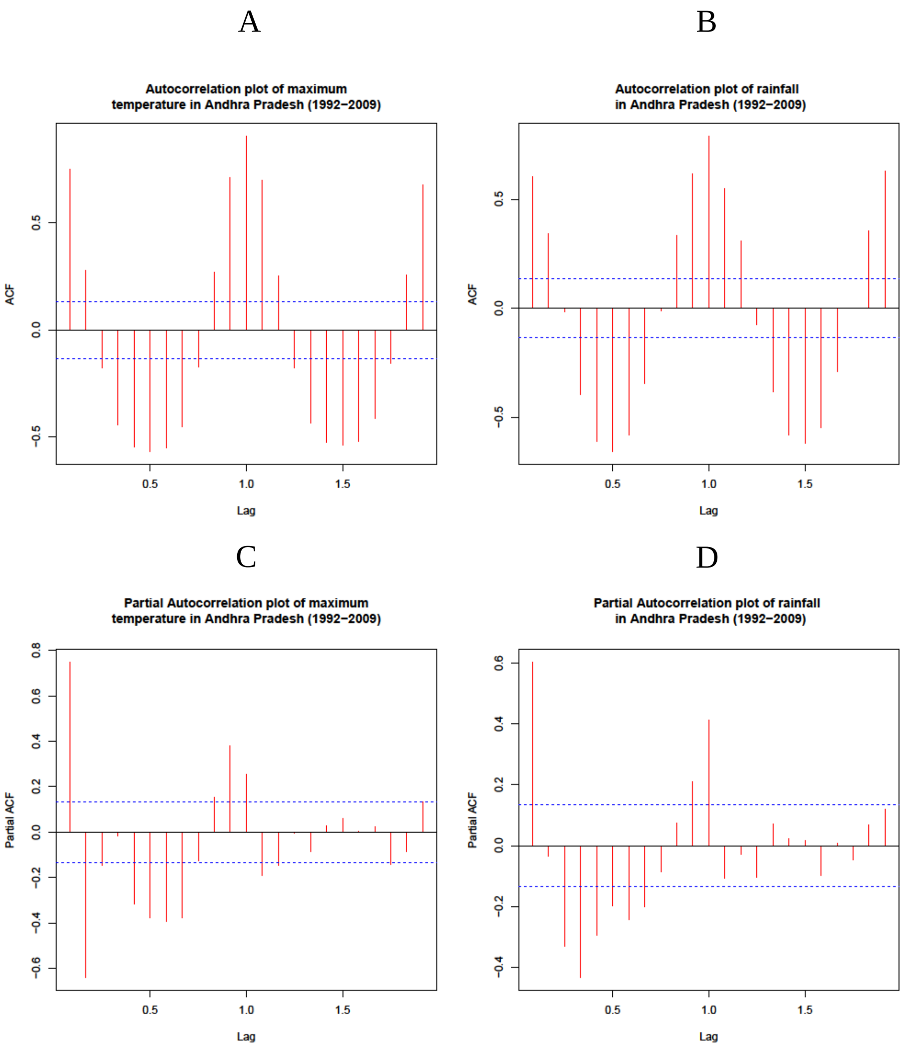


Figure 3.4: Plots of (A) Autocorrelation function (B) Partial autocorrelation function of maximum temperature and rainfall. The X-axis gives the number of lags in years and, the y-axis gives the value of the correlation between -1 and 1. Blue dashed lines indicate the 95% confidence intervals, within which the correlation is non-significant.

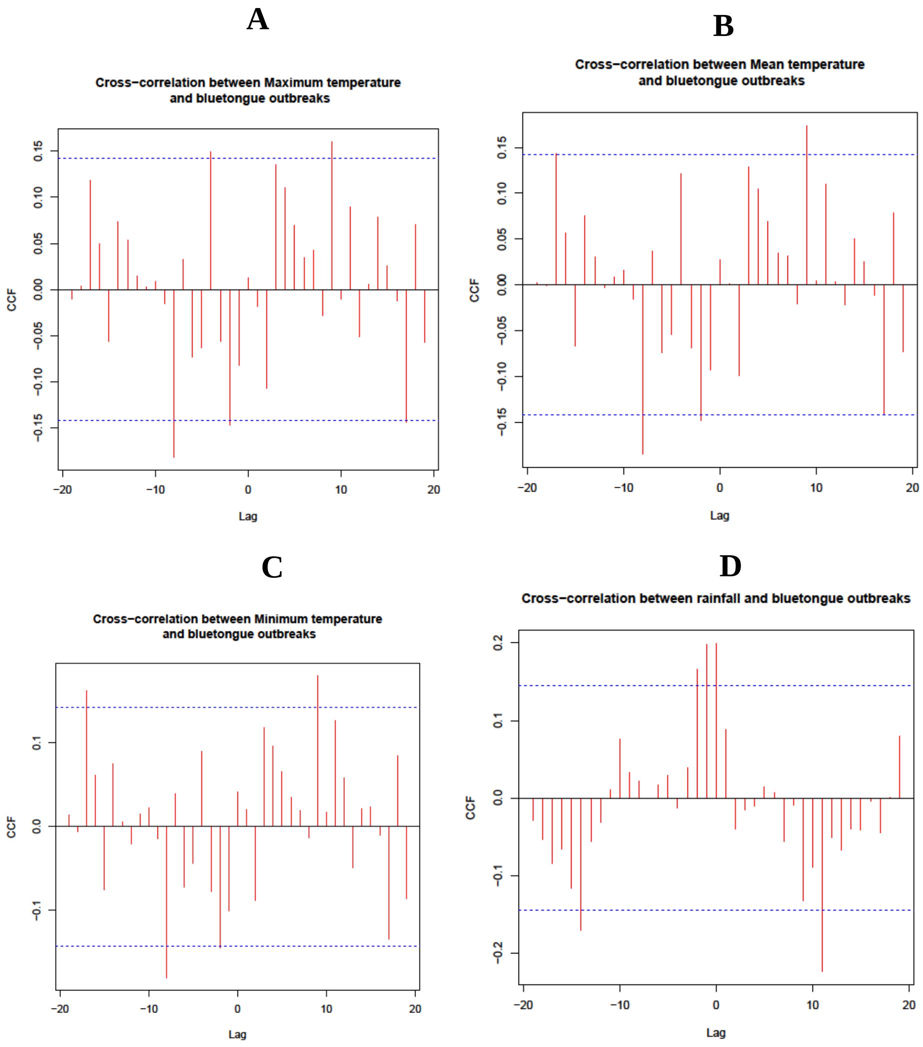


Figure 3.5: Cross-Correlation functions of bluetongue outbreaks with rainfall, maximum, mean and minimum temperatures. (A–D) Cross-correlation functions (CCF) with (A) Maximum temperature (B) Mean temperature, (C) Minimum temperature and (D) Rainfall. The blue dashed lines are the 95% confidence intervals for the cross-correlation between two series that are white noise.

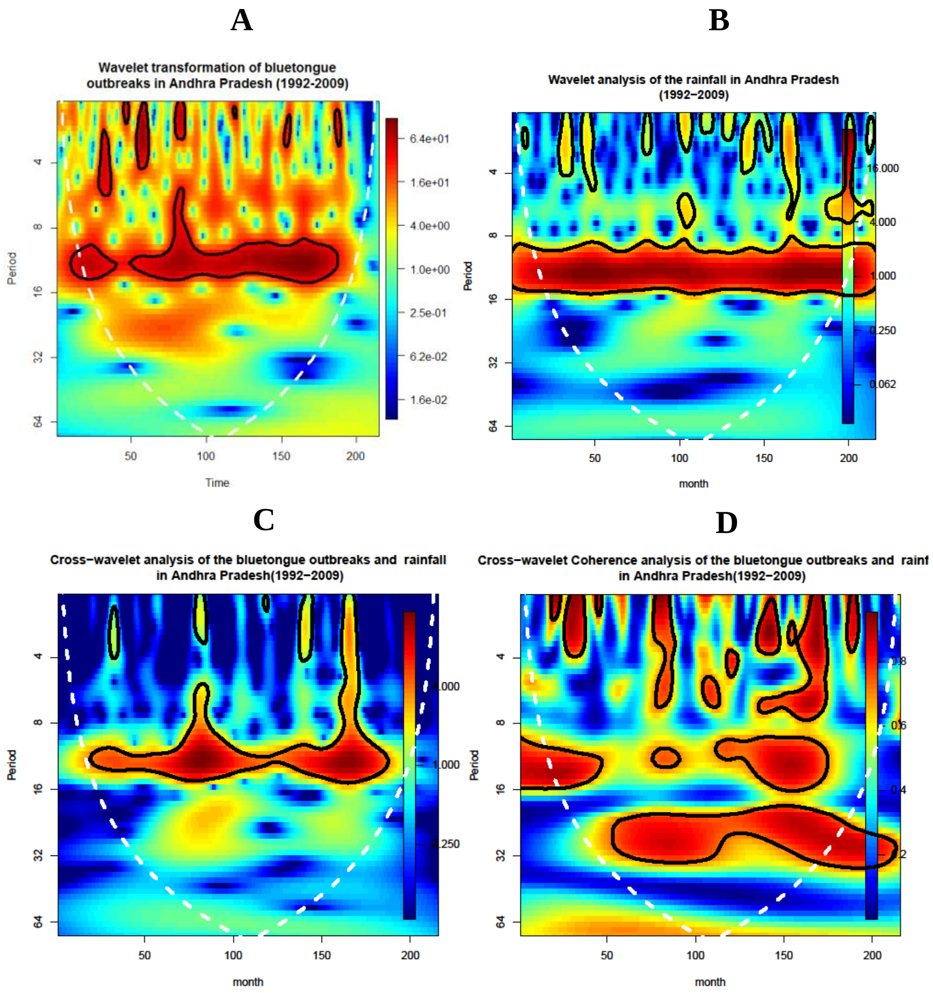


Figure 3.6: Dominant frequencies in the monthly bluetongue outbreaks time series & monthly rainfall time series and their cross-correlation. Wavelet power spectrum- The white dotted line is the cone of influence indicating the region of time and frequency where the results are not influenced by the edges of the data and are therefore reliable. The solid black line corresponds to the 95% confidence interval and the areas within this black solid line indicate significant variability at the corresponding periods and times. (A) & (B) wavelet power spectrum of the bluetongue outbreak series and rainfall respectively. The wavelet spectrum is shown with power increasing from blue to red colours. (C) Cross-wavelet power spectrum between bluetongue outbreaks and rainfall (D) Cross-wavelet coherence (correlation) between the two time series. Spectrum power in (C) and coherence in (D) increases from blue to red (for (D) on the scale of 0 to 1).). X-axis: time in months from January 1992 (= month 1), Y-axis: localised periodicity in months.

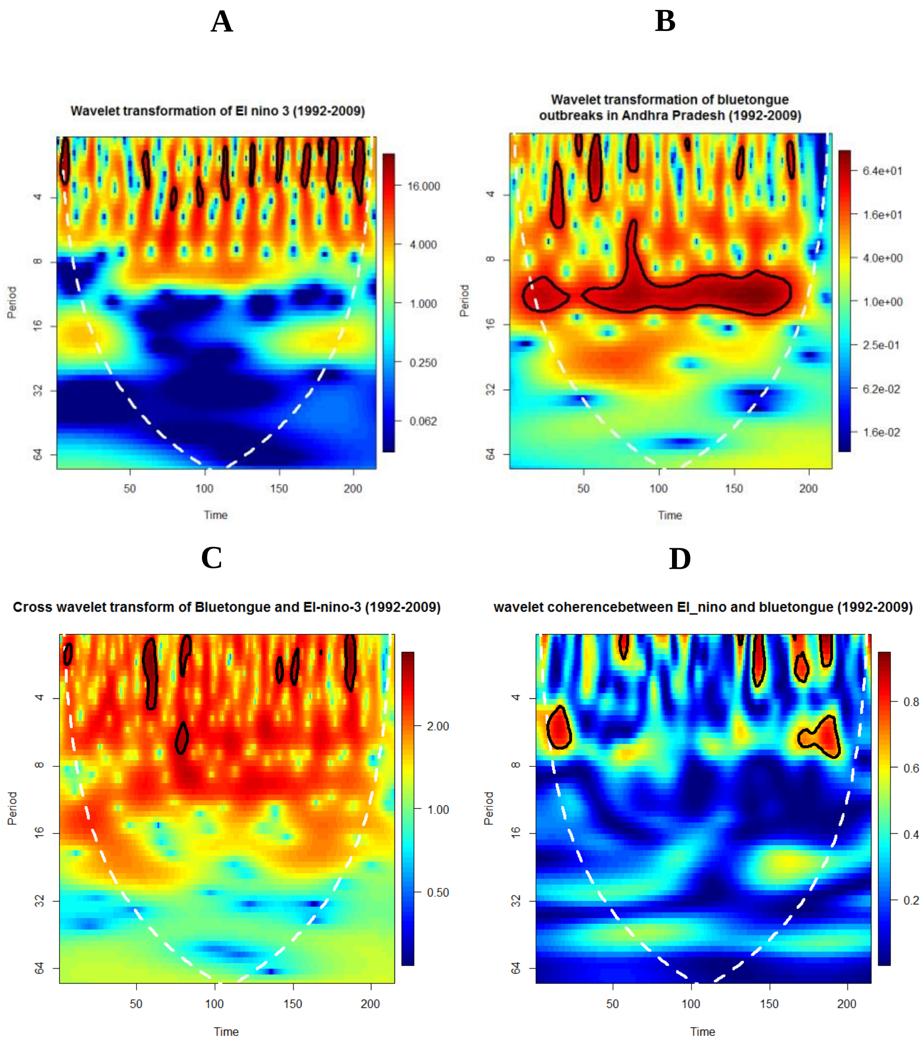


Figure 3.7: Dominant intra-annual frequencies in the monthly bluetongue outbreaks time series & monthly El Niño 3 time series and their cross-correlation. Wavelet power spectrum- The white dotted line is the cone of influence indicating the region of time and frequency where the results are not influenced by the edges of the data and are therefore reliable. The solid black line corresponds to the 95% confidence interval and the areas within this black solid line indicate significant variability at the corresponding periods and times. (A) & (B) wavelet power spectrum of the bluetongue outbreak series and rainfall respectively. The wavelet spectrum is shown with power increasing from blue to red colours. (C) Cross-wavelet power spectrum between bluetongue outbreaks and rainfall (D) Cross-wavelet coherence (correlation) between the two time series. Spectrum power in (C) and coherence in (D) increases from blue to red (for (D) on the scale of 0 to 1).). X-axis: time in months from January 1992 (= month 1), Y-axis: localised periodicity in months.

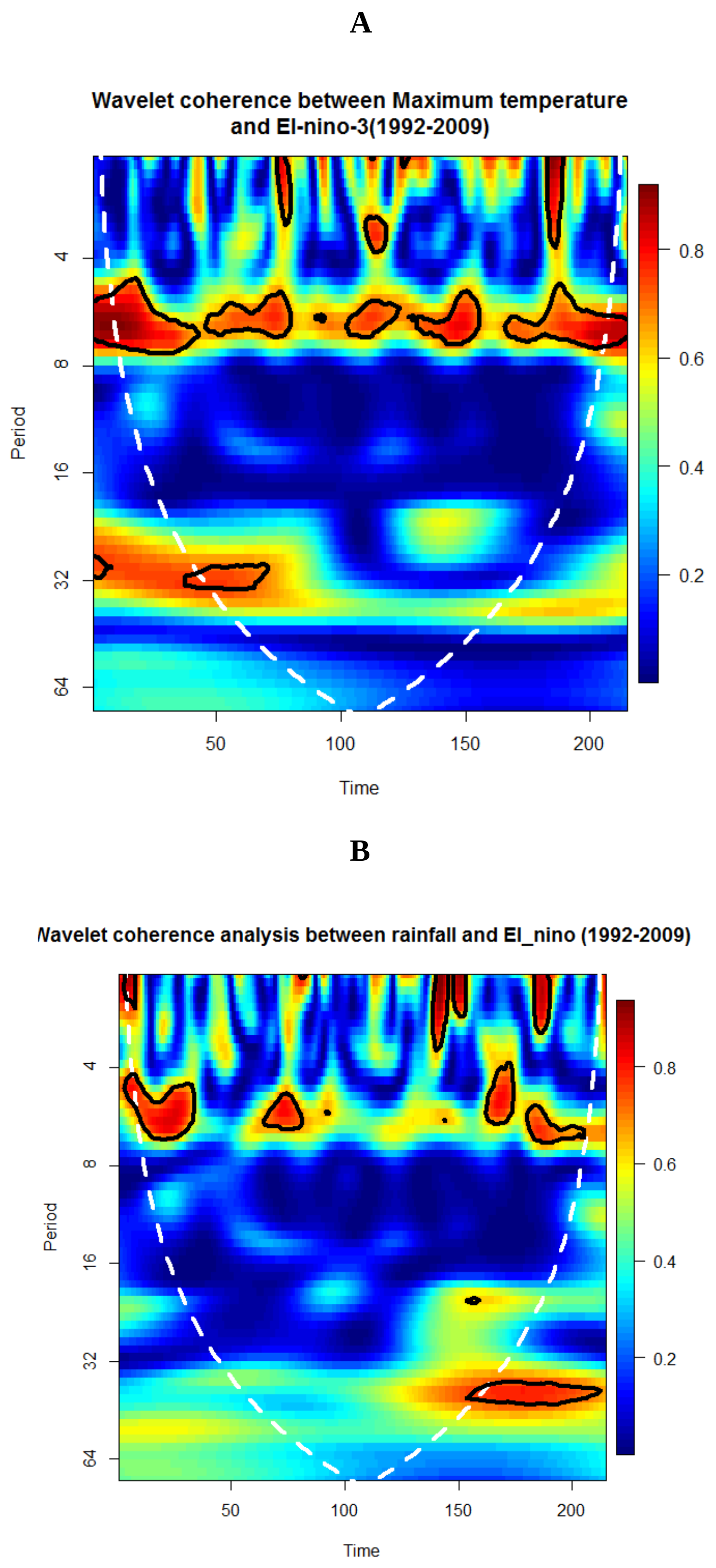


Figure 3.8: correlation between El Niño 3 with maximum temperature (A) and rainfall (B) Cross-wavelet coherence (correlation) and the wavelet spectrum is shown with coherence increasing from blue to red colours(scale is from 0 to 1.0). X-axis: time in months from January 1992 (= month 1), Y-axis: localised periodicity in months.

No.	**Variables**	**DIC**
1.	**Best combination of monthly rainfall variables**	
2.	Rainfall at lag 2 + AR(1)	**494.072**
3.	Rainfall at lag 0 & 2 + AR (1)	494.406
4.	Rainfall at lag 1& 2 + AR (1)	494.585
5.	Rainfall at lag 0,1 & 2 + AR (1)	494.848
6.	Rainfall at lag 0 + AR (1)	500.888
7.	Rainfall at lag 0 & 1 + AR (1)	502.821
8.	Rainfall at lag 1 + AR (1)	503.229
9.	Rainfall at lag 0 + AR (1)	500.888
9.	**Best combination of monthly temperature variables**	
10.	Maximum temperature at 2, 4 & 8 lag + AR (1)	**499.299**
11.	Maximum temperature at 4 & 8 lag + AR (1)	499.616
12.	Maximum temperature at 2& 4 lag + AR (1)	501.215
13.	Maximum temperature at 4 lag + AR (1)	502.686
14.	Maximum temperature at 2 + AR (1)	503.345
15.	Maximum temperature at 8 lag + AR (1)	503.413
16.	Maximum temperature at 2 & 8 lag + AR (1)	503.96
17.	**Combination of best monthly rainfall and monthly temperature**	
18.	Maximum temperature at 2, 4 & 8 lag and rainfall at lag 2 + AR (1)	**498.601**
19.	Maximum temperature at 2& 4 lag and rainfall at lag 1 + AR (1)	499.241
20.	Maximum temperature at 2& 4 lag and rainfall at lag 2 + AR (1)	499.241
21.	Maximum temperature at lag 4 and rainfall at lag 2 + AR (1)	502
22.	Maximum temperature at 2& 8 lag and rainfall at lag 2 + AR (1)	502.62
23.	Maximum temperature at 8 lag and rainfall at lag 2 + AR (1)	503.11
24.	Maximum temperature at 2 lag and rainfall at lag 2 + AR (1)	503.913

Table 3.1: Selection of monthly models (rainfall and temperature) and combination of rainfall and temperature variables based on reduced DIC.

Variable	**Mean (sd)**	**Credible interval**
Best model		
Intercept	-2.509 (0.494)	-3.537, -1.577
Rainfall at lag 2	0.026 (0.003)	0.018, 0.033
Temporal autocorrelation at lag1	0.5274 (0.4051)	

Table 3.2: Mean co-efficient and credible interval of the best model (rainfall at lag 2, plus AR (1))

Model	**RMSE**	**DIC**
AR(1) model	0.32	497.08
Rainfall at lag 2 + AR(1) model	0.38	494.07
Rainfall at lag 2 model	57.51	6985.8

Table 3.3: Root mean square error (RMSE) and DIC for rainfall at lag 2 with AR (1), rainfall at lag 2 and AR (1) only model. The total variance of the raw data was 3492

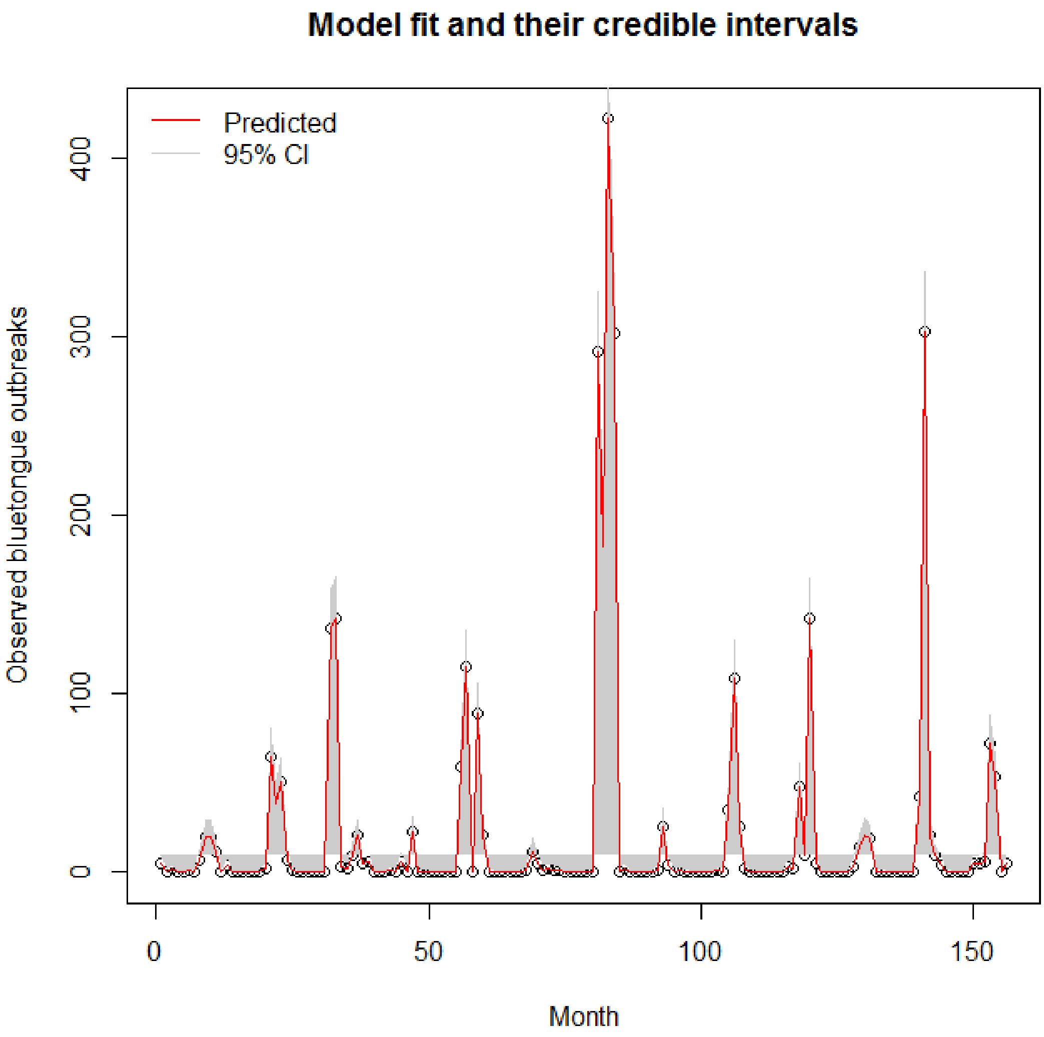


Figure 3.9: Plot of best model fit with rainfall at lag 2 and the AR(1) error term. Grey areas correspond to the 95% credible interval, red solid line is the predicted number of outbreaks and the circles are the observed number outbreaks.

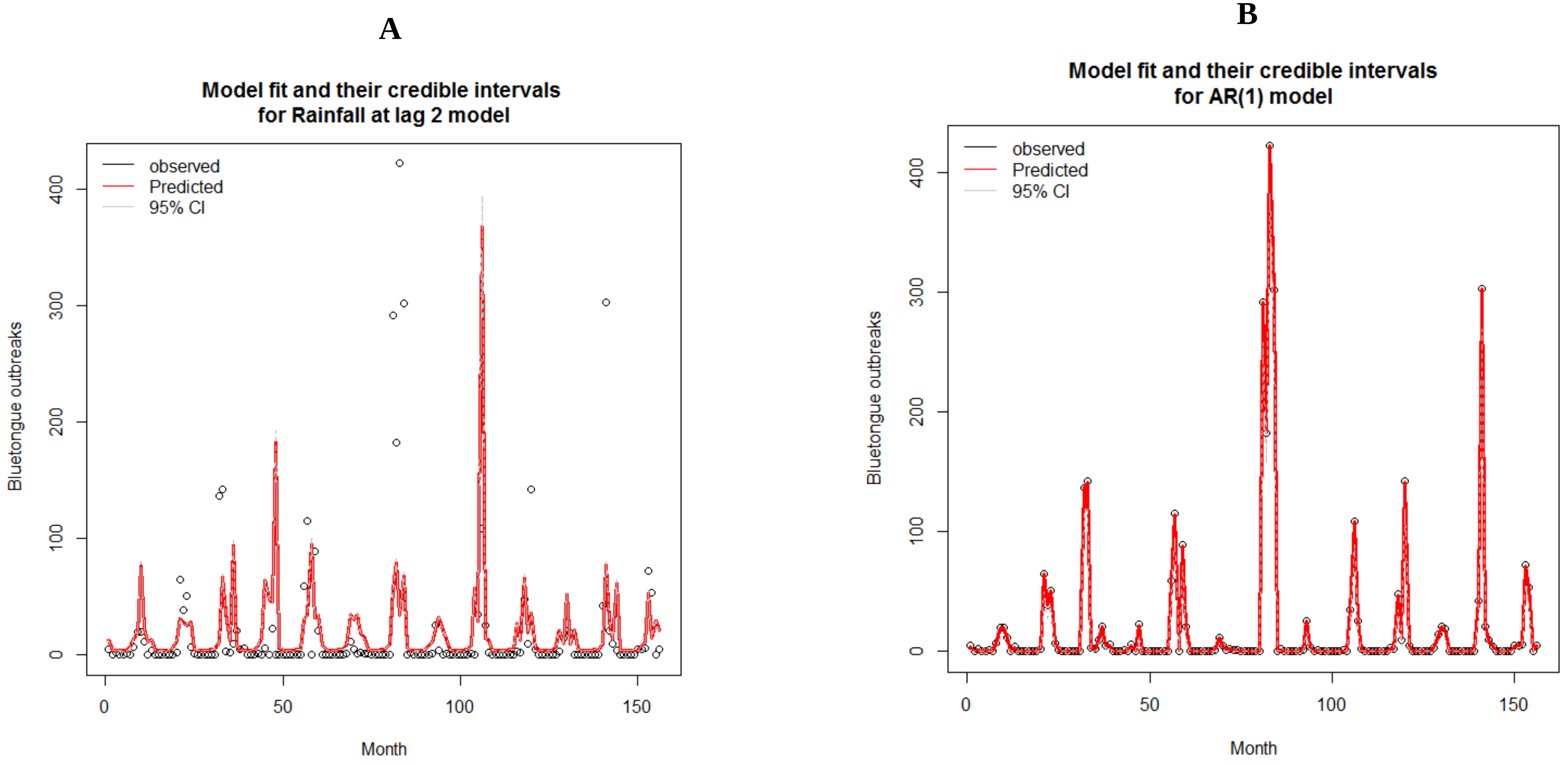


Figure 3.10: Plot of model fits with A) rainfall at lag 2 only (no AR(1)) and B) AR(1) only (no rainfall at lag 2).Grey lines correspond to the 95% credible interval, red solid line is the predicted number of outbreaks and the black circles are the observed number outbreaks.

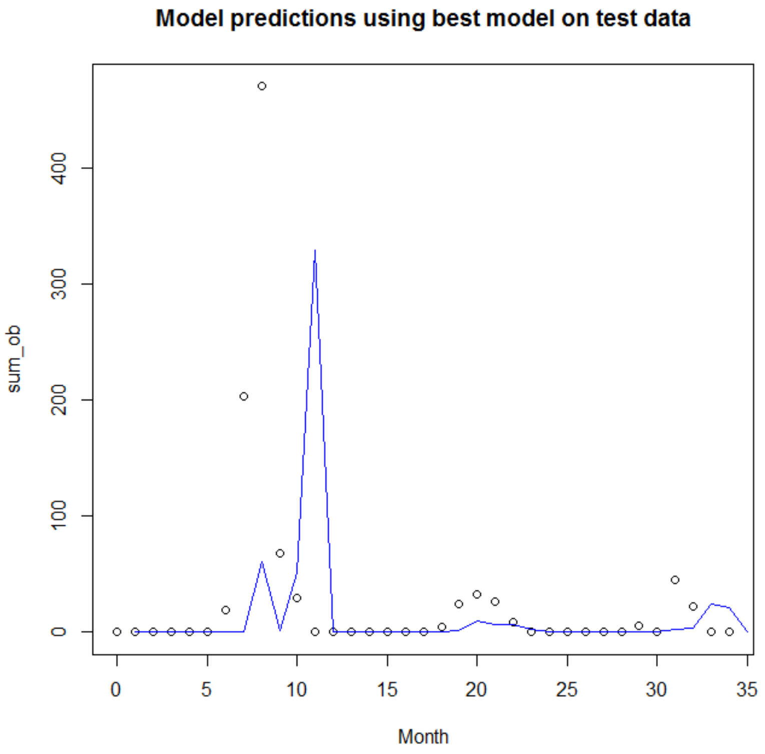


Figure 3.11: Plot of observed bluetongue outbreaks and predictions on "out of fit" data with best model with lowest DIC (Rainfall at lag 2 with AR (1)).(Black open circles: observed bluetongue outbreaks, blue line: predicted bluetongue outbreaks)

3.4 Discussion

The present study considered the effects of climate on monthly bluetongue outbreaks in Andhra Pradesh by accounting for temporal autocorrelation using a Bayesian Poisson regression approach. Numbers of bluetongue outbreaks increased under conditions of high rainfall two months previously. The lagged effect of satellite derived temperature variables (negatively correlated) and NDVI (positively correlated) has been shown to be significant in the temporal epidemiology of bluetongue in Israel

(Purse, et al., 2004). The semi-aquatic conditions created by high rainfall are favourable for midge breeding and there are reports of correlations between abundance of *C.imicola* and sero-conversions in domestic animals after the start of monsoon season in Tamil Nadu state (Udupa 2001).

Overall, the best model was also one of the most parsimonious ones, including only rainfall at lag 2 and temporal autocorrelation in the error term; this model was therefore used for making prediction on the 'out-of-fit' data.

Wavelet analysis is advocated along with cross-correlation analysis to identify periodicities and significant lags in vector borne diseases (Cazelles et al., 2007). This study detected only an annual cycle in the cases of bluetongue and the rainfall time series, possibly due to the relatively short time series involved . In general, a time series has to be at least six times as long as the longest period within it that may be demonstrated statistically (Chatfield, 2013). Thus the 17 year time series here would not be expected to reveal significant periodicities of longer than about two to three years.

Wavelet coherence analysis which measures the strength of correlation at specific times and periodicities helps to understand the relationship between two time series (Cazelles et al., 2008). The wavelet coherence graph identified a high and relatively persistent correlation of bluetongue and rainfall at periodicities of between two and three years (Fig. 3.6D), although the strength of both signals at these periods appears to be relatively weak (Fig. 3.6A and Fig. 3.6B).

There was no correlation between the bluetongue outbreaks and El-Niño, except for two brief periods of correlation at a periodicity of six months. This absence of any strong, consistent relationship between BTV outbreaks and El Niño rules out any possibility of using the latter to predict the former at the present time. Longer time series are required to investigate what links, if any, occur between BTV outbreaks and any El-Nino associated phenomenon (including climate variables). Ideally these links should involve periodicities of > 1 year (since within-year periodicities are determined by more regular seasonal events) so that future disease forecasting may be on a longer term basis, if this is at all possible.

The best model was able to differentiate between years with a larger number of outbreaks and years with fewer outbreaks (Fig. 3.11), although the sample size (n=3) was very small to detect inter-annual variability. The model also did not predict seasonal timing of the outbreaks particularly well.

The model with only a single climatic variable (rainfall at lag 2) was the most parsimonious model with temporal autocorrelation. The results demonstrated that the intrinsic factors (models with temporal autocorrelation) dominate the extrinsic factors (model with covariate) considering the huge drop in DIC (6985 units). The domination of intrinsic factors over the extrinsic factors is important with respect to BTV virus control strategies and future surveillance activities. There is need for regular surveillance of vectors, serotype distribution in different seasons and also the immune status (herd immunity data) to better understand the significant and dominant effect of intrinsic mechanisms in this analysis.

Nevertheless, the combined model (with covariate and temporal autocorrelation) can be used in very approximate, qualitative forecasting of BTV outbreaks, and urgently needs further improvement by investigating the nature of the important AR (1) effect.

Overall the results highlight the importance of lagged climatic effects on bluetongue outbreaks. The selection of rainfall suggests the importance of these conditions on the breeding of midges and subsequent transmission of the virus, but the effect of climate is more on seasonality of BTV outbreaks. Although long-term periodicity was not identified in either the bluetongue series or the rainfall series when analysed with wavelet or cross wavelet analysis, there was strong correlation of the dominant frequencies at two and three year periods in wavelet coherence analysis. Wavelet analysis has been previously used in epidemiological time series (Chaves & Pascual, 2006; Onozuka, 2014). There was no significant long term periodicity in the bluetongue outbreaks data and also on the residuals after fitting a Poisson model with autoregressive error structure

The performance of the AR (1) alone (Table 3.3 and Fig 3.10 B) was better than the model with rainfall at lag 2 and this section discusses its significance. The difference between purely autoregressive models (inclusion of the number of previous outbreaks or cases) and models which include AR (1) in the error term (which substitutes for the past number of cases or outbreaks) in the frequentist domain was discussed in Chapter 1. The advantage of specifying an AR(1) relationship in the error terms of a Bayesian model is that each component (meteorological variables and autoregressive component) are derived from a prior distribution and the posterior distribution is used for drawing inference and thus the role of different components can be estimated. There are purely autoregressive models for Poisson data in the frequentist domain (i.e. the autoregressive element is not in the error term), but this makes it difficult to quantify the role of the different components in the model. Quantifying the variance explained by different components is not straightforward in Bayesian framework also, but the contribution of different components (fixed and random effect) can be compared by change in DIC by fitting fixed effect and random effect models separately. The dominance of AR (1) term in the best model over rainfall at lag 2 is interesting considering the role of other variables (unmeasured variables) which the AR (1) term captures. The other unmeasured variables can again be either extrinsic or intrinsic or both. The unmeasured extrinsic variable may be relative humidity, wind speed, soil moisture or some other variable of importance to the life cycle of the vectors or to the transmission of BTV. The role of herd immunity as one of the intrinsic factor was discussed earlier. What was particularly interesting here was that whilst the PACF for the BT case numbers shows no effects beyond a lag of 1 year (Fig 3.3B) the equivalent PACF of the climate variables did so (Fig. 3.4D). This rather perplexing result indicates that whilst the climatic system seems to have a ‘memory’ lasting longer than 1 year, the BT system does not. It is perhaps possible that the inter-year memory effects of herd immunity are obscured in a vector-borne disease such as BT because the numbers of vectors produced each year is highly variable, so the ratio of infected vectors to hosts may be relatively independent of the number of past BT cases.

Chapter 4

Understanding spatial variation in occurrence of bluetongue outbreaks between districts in South India using a Bayesian Poisson regression model

4.1 Introduction

Midge-borne disease systems are ecologically complex, with transmission in India and elsewhere often involving several ruminant hosts and biting midge (*Culicoides*, Diptera:Ceratopognidae) vector species, diverse landscape and environmental conditions within a single region. Though monsoons are thought to govern the size and timing of epidemics in India (Prasad et al., 2009), the severity of the disease varies substantially between districts, even within areas subject to similar monsoon conditions, suggesting that other landscape and host variables also affect transmission. More than 22 BTV serotypes are circulating in the country, from distinct geographic origins (Maan et al., 2012; Sreenivasulu et al., 2003). There is a paucity of systematic studies of Indian midge distributions and vectorial capacities (Ilango, 2006), but both dung-breeding (e.g. *C. oxystoma*, which breeds in buffalo dung) and moist soil breeding midges (e.g. *C. imicola, C. schultzei, C. peregrinus*), have been found to be abundant in BT-affected districts in different states (Reddy & Hafeez, 2008). Mixed farming by the small and marginal farmers of South India often involves indigenous cattle maintained for draft purpose and buffalo for milk production. Variation in the mixture of different livestock species kept in the different regions of India is expected to affect the impact of midge-borne BTV in South India. This chapter investigates the roles of hosts, climate and landcover in determining the severity of bluetongue outbreaks across South India by employing an All Subset Method (ASM) for variable selection in a Bayesian framework, also accounting for spatial autocorrelation.

4.1.1 Diversity of climate, hosts, and landcover on BTV outbreaks and vector distribution

In many areas affected by BT worldwide, disease impacts have been linked to rainfall patterns (Baylis et al., 1999; Walker, 1977). Annual rainfall in South India is influenced by the South-West monsoon system (June-september) and the North-East monsoon system (October- December). Andhra Pradesh, Karnataka and some parts of Tamil Nadu receive South-West monsoon rainfall (Fig. 4.1C), while the North-East monsoon rainfall (Fig. 4.1B) covers coastal Andhra Pradesh, some parts of Karnataka and most of Tamil Nadu. *Culicoides imicola* breeds in wet soil enriched with organic matter (Meiswinkel et al., 1994) and its abundance is related to temperature and rainfall. Temperature plays a significant role in the survival and fecundity of *Culicoides* (Mellor, 2000) and precipitation can provide semi-aquatic habitats for the development of larvae.

India is home to 74 million sheep, 6.8% of the world sheep population (FAOSTAT 2010) and 12.71% of the total livestock population of the country (Livestock census 2012). There are more than 40 sheep breeds in India, of which 14 are present in South India. Worldwide, certain breeds of sheep and wild ruminant species (e.g. white-tailed deer) are highly susceptible to BT disease (Maclachlan et al., 2009). Sheep that are native to tropical and subtropical regions of the world where BTV is enzootic are usually resistant to BT, whereas fine-wool European breeds such as the Merino are highly susceptible. The same situation exists in India, where past outbreaks of BT have been detected not only in exotic breeds (including Rambouillet and Merino) but also in crossbreeds of sheep, whilst local breeds are considered to be relatively resistant (Prasad et al., 2009) (Lonkar et al., 1983), though quantitative genetic breed susceptibility studies are lacking. In South India, Nellore, Ramnad white and Trichy black are thought to be more susceptible to BTV than other breeds (Rao et al., 2014)

Although host species and type are very important in bluetongue epidemiology, very few studies have examined their roles (Baylis et al., 2004, Witmann et al., 2001) in determining the distribution of disease patterns. A recent study (Acevedo et al., 2010) to predict spatial patterns in *C.imicola* abundance found that host abundance explained the maximum variance of all the predictors considered. In the past there were reports of high seroprevalence of BTV virus in livestock species other than sheep, such as cattle and buffalo, without clinical disease (Prasad et al., 2009).

Particular land covers may favour the presence and abundance of potential vectors for bluetongue or provide suitable habitat for grazing of susceptible hosts. For example, forest and pasture areas were found to increase the risk of BTV-8 spread in North-Western Europe (Faes et al., 2013) and this was attributed to these being preferred habitats for the key European vector species group, the *C. obsoletus* group. *Culicoides imicola* preferred sparsely vegetated areas, whilst species in the *Obsoletus* group favoured shaded habitat (Conte et al., 2007). The influence of land use has less often been considered in tropical and subtropical countries but in India it is expected that irrigated and rain-fed agricultural areas may be more likely to contain suitable semi-aquatic breeding sites for midges, whilst forested, high altitude areas and urban areas will be unsuitable for grazing.

4.1.2 Spatial autocorrelation

Understanding how geographical variability in hosts, vectors and climate interact to produce variation in disease severity across districts requires quantitative methods that can deal with collinearity and spatial dependency in errors, which may arise due to intrinsic processes (disease spread between districts) and extrinsic effects (arising from spatial autocorrelation of environmental variables). Collinearity occurs when predictor variables are highly correlated, hampering discrimination of their individual effects on the outcome or dependent variable. In disease epidemiology, observations which are nearer to each other have errors that are more similar than observations farther apart. Such spatial autocorrelation inflates model accuracy but also the estimated explanatory power of environmental predictors (Dormann et al., 2007). Spatial autocorrelation can also be a problem if certain independent variables (with spatial structure) are unavailable and therefore have to be omitted.

In generalized linear models, all parameters are modelled as fixed effects and estimated by Maximum Likelihood (ML) methods. In ecological studies, when there is spatial dependence, and key covariates may be missing, the ML approach often leads to unsatisfactory estimates of the district level risk due to extra-Poisson variation (Clayton, 1996)

Bayesian generalized linear mixed models (Breslow & Clayton, 1993) overcome these problems by explicitly modelling the missing covariates and spatial dependence as random effects, through a prior distribution (Clayton, 1996). The Besag-York-Mollie (BYM) model (Besag et al.,

1991) can account for spatially structured and spatially unstructured random error variation (the latter arising from unmeasured non-spatial predictors) as well as fixed effects of environmental predictors. Bayesian disease mapping has been developed for many diseases (Banerjee et al., 2004; Besag et al., 1991) using MCMC (Markov Chain Monte Carlo) algorithms for parameter estimation for both chronic non-infectious diseases as well as for vector borne diseases (Alexander et al., 2000; Diggle, et al., 2002b). Recently, an approximate method for parameter estimation in Bayesian frameworks has been proposed (Rue et al., 2009). This uses integrated nested Laplace approximations (INLAs) to estimate the posterior marginals of interest and can be computed easily with vastly reduced computation time (compared to MCMC) in R (R Development Core Team, 2005) using the INLA library (Martino & Rue, 2009). The computatonal efiiciency of INLA makes possible the fitting of all possible combinations of variables (All Subsets Method) in a Bayesian framework. Using this BYM model approach fitted in INLA (Blangiardo et al., 2013) the present chapter investigates the role of climate, land-cover and availability of livestock hosts in explaining geographical variation in the severity of BT outbreaks across districts in South India with the following hypotheses in mind:

1. The severity of outbreaks will be greater in areas with greater availability of land-cover-types containing water bodies or irrigated areas that provide breeding habitat for *Culicoides* (Diptera: Ceratopogonidae) midges;

2. Outbreaks will be fewer or absent in areas with closed forest because farming of any sort usually does not occur in closed forests but may occur in open forest types that are used by farmers for grazing;
3. The severity of outbreaks will increase as sheep numbers increase, particularly as the numbers of certain local and exotic breeds increase;
4. The severity of outbreaks will increase as cattle and buffalo numbers increase, since virus may circulate silently in these reservoir hosts and increase levels of disease in co-occurring susceptible hosts

The three states of South India (Andhra Pradesh, Karnataka and Tamil Nadu) are distinct not only in terms of their geography, land use pattern, climate and host and breed diversity but also in their disease reporting systems and veterinary expertise (veterinary colleges, disease diagnostic laboratories, number of veterinary hospital). Therefore data quality varies from state to state. Given the differences in both the disease systems and the reporting in the different states, modelling was carried out separately for the three states, and compared with a combined, South India model.

4.2 Materials and Methods

4.2.1 Bluetongue outbreak data

District level (admin-2) monthly BT outbreak data (1992-2009) were provided by PD_ADMAS (Project Directorate on Animal Disease Monitoring and Surveillance) which maintains the livestock diseases database for India and collates outbreak data every month from different sources. The analysis was restricted to data from three states of South India namely Karnataka (n = 27 districts), Andhra Pradesh (n = 23 districts) and

Tamil Nadu (n = 30 districts), in which outbreaks are regularly reported, with 61 out of the total of 80 districts having reported outbreaks of BT at one time or another over the 18 year study period. Despite the fact that there is high sero-prevalence of BT antibodies in most of the states of India (Bandhyopadhyay & Mallick, 1983; Kakker et al., 2002; Bhanuprakash et al., 2007), there are infrequent clinical outbreaks outside these three southern states. The mean annual number of outbreaks per district over the study period was calculated and was used as the dependent variable in all the models described here.

4.2.2 Land-cover data

The proportions of each district covered by ten land-cover classes were extracted from the Global cover land-cover map (GlobCover, Defourny et al., 2006) using the Zonal Statistics option in ArcMap 10.1 (ESRI, Inc., Redlands, CA, U.S.A.). These were logit-transformed (as they were not normally distributed) and each considered individually as predictors in the analysis. The ten classes were selected from the original 24 available in GlobCover, due to their assumed importance for BT epidemiology, either because they were likely to contain favourable semi-aquatic breeding habitats for *Culicoides* vectors (post-flooded/ irrigated cropland, rain-fed and mosaic croplands, water bodies, classes 1 to 3) or because they were likely to be favourable (closed to open forest, classes 4, 5 and 7) or unfavourable for grazing of livestock (closed forest, including different percentages of cover, classes 6 and 8). Urban areas (class 9) and water bodies (class 10) are areas with low livestock density and were expected to be unfavourable for disease occurrence.

4.2.3 Host species and sheep breed data

Densities of host species, including indigenous sheep, non-descript sheep, crossbred & exotic sheep, goats, crossbred cattle and buffaloes, were extracted from the database of National Livestock census data (http://www.dahd.nic.in/) and log-transformed because the absolute values were not normally distributed. Out of over 40 breeds of sheep present in India (Patnayak, 1988), only 14 breeds are present in South India. The indigenous breeds are; Bellari, Mandya, Deccani, Hassan, Nellore, Coimbatore, Kengur, Kilakarsal, Madres Red, Mercheri. All the exotic and crossbred Sheep are grouped into one category referred to here as 'exotic & crossbred Sheep'. Finally, the single 'non-descript sheep' category is of all the indigenous breeds which cannot be identified with any certainty or do not have more than 50% similarity to any recognised breed. The distributions of several host types in South India are shown in Fig. 4.3.

4.2.4 Rainfall and temperature data

Monthly Rainfall Estimates (RFE) were obtained from the NOAA/Climate Prediction centre RFE 2.0. (Xie et al., 2002). Seven year averaged values (2004-2010) of different monsoon rainfall (South-West and North-East) and annual rainfall for districts were extracted using the Zonal Statistics function in ArcMap 10.1 (ESRI, Inc., Redlands, CA, U.S.A.). The South-West monsoon occurs from June to September and the North –East monsoon from October to December. Thus the monsoon rainfall variables were calculated as the sums of the monthly rainfalls for these respective periods. The annual mean temperature layer (1950-2000) at 1km spatial resolution was obtained from Worldclim (Hijmans et al., 2005).

Thus for the present analysis average annual values of BTV outbreaks per district for the period 1992-2009 were related to a single time point estimate of land-cover (2006 map), host abundance (2007 census) and mean seasonal rainfall data (2004-2010) and a single time point estimate of mean annual temperature (derived from a single climate surface based on data from 1950-2000).

4.2.5 Modelling approach

Relationships between the average annual number of BTV outbreaks (Y) and environmental predictors were quantified using a generalised linear mixed model with Poisson errors, implemented in a Bayesian framework. The probability function for Y is demonstrated in Eq. (1).

$$
\begin{aligned}
&\Pr\left(\mathrm{Y} = \mathrm{y}/\mu\right) = Poisson(\mu_i) \qquad (1)\\
&\log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_n x_{ni} + \upsilon_i + \nu_i\\
&\upsilon_i \mid \upsilon_{j \neq 1} \sim Normal\left(m_i, s^2{}_i\right)\\
&m_j = \frac{\sum_{j \in N(i)} \upsilon_j}{\neq N_{(i)}}\\
&s_i^2 = \frac{\sigma_\upsilon^2}{\neq N_{(i)}}
\end{aligned}
$$

OR

$$
\begin{aligned}
&\Pr\left(\mathrm{Y} = \mathrm{y}/\mu\right) = Poisson(\mu_i) \qquad (2)\\
&\log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_n x_{ni} + \upsilon_i
\end{aligned}
$$

OR

$$
\begin{aligned}
&\Pr\left(\mathrm{Y} = \mathrm{y}/\mu\right) = Poisson(\mu_i) \qquad (3)\\
&\log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_n x_{ni} + \nu_i
\end{aligned}
$$

OR

$$
\begin{aligned}
&\Pr\left(\mathrm{Y} = \mathrm{y}/\mu\right) = Poisson(\mu_i) \qquad (4)\\
&\log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_n x_{ni}
\end{aligned}
$$

Where β_0 is the intercept β_1 to β_n be the coefficients for the fixed effects of predictors.

(x_{1i} to x_{ni}) in each district i. υ_i is a structured spatial component assuming Besag-York-Mollie (BYM) specification (Besag et al., 1991), modelled using an intrinsic conditional autoregressive structure (iCAR).

$\neq N_{(i)}$ is the number of districts that share boundaries with the *i*-th district, and v_i is the unstructured spatial effect in each district, modelled using an exchangeable prior v_i ~ Normal (0, $\sigma_{\upsilon}{}^2$). iCAR is based on a set of districts that share boundaries for which an adjacency matrix is defined, listing for each district all other districts with which it shares a boundary or adjacency. Weights are defined for those adjacencies, and have a value of 1 when two districts share a boundary and a value of zero when they do not.

4.2.6 Model building and selection of predictors

Variable selection is critical to understanding the importance of the impact of individual environmental predictors upon the distribution of BT outbreaks, and it can be undertaken using a wide range of frequentist and Bayesian approaches (Efron et al., 2004; George, 2000; Miller, 2002). Given the computational efficiency that INLA offers over MCMC methods, it was possible to implement a modified All Subsets approach to variable selection, where all possible combinations of the total *p* explanatory predictors, from size 1 to *p*, were fitted (making 2^p-1 combinations in all) and the most parsimonious model was selected using information criteria. Pairwise Pearson correlation analyses were performed on all 32 predictors to identify pairs of predictors that were highly correlated (r > 0.7, p < 0.001) (leading to the removal of 8 breed and host predictors, and one land-cover class). The two correlated variables were removed by fitting a univariate model for each of the correlated variables and the model with lower DIC was retained. In the South India model there were 25 predictors (9 land-cover, 4 climate, and 12 host predictors). In Andhra Pradesh there were 22 variables (9 land cover, 4 climate and 9 host variables), in Karnataka there were 25 variables (9 land cover, 4 climate and 12 host variables) and in Tamil Nadu there were 28 variables (9 land cover, 4 climate and 15 host variables) considered for selection by the models (Tables 4.1 & 4.2). Since it was impossible even in INLA to fit all possible combinations for the entire predictor dataset, the all subset approach was applied first to each predictor variable set alone (i.e. land-cover or climate or host type). This approach identified the best land-cover, climate and host models, and this was done for each state in turn and then for all states together (covering the whole of South India). For each geographical area, the best model (Eq 1) was identified within each category as the model with the lowest Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002). DIC is a generalisation of the Akaike Information Criterion (AIC), and is derived as the mean deviance adjusted for the estimated number of parameters in the model, striking a balance between model fit and complexity. This approach allows a measure of out-of-sample predictive error and prevents over-fitting (Gelman & Hill, 2006). Once the best model in each category had been identified for each geographical area, all possible model combinations of the constituent predictors were fitted and the best combined model was again identified using the DIC.

Once the best model was identified in each category of the predictor variables by including both structured and unstructured heterogeneity (Eq 1) and the covariates were analysed by fitting four different models: BYM model with covariates (equation 1); a Besag model with covariates(equation 2); an *i.i.d* model with covariates (equation 3); and finally a covariate only model (equation 4). These four types of model were fitted in order to understand the role of the different components in explaining bluetongue outbreaks in South India.

The proportion of variance explained by the spatially structured component and the unstructured component within a BYM model was estimated by equation 5:

$$S_{\upsilon}{}^{2} = \frac{\sum_{i=1}^{n}(\upsilon_i - \overline{\upsilon})^{2}}{n-1} \tag{5}$$

$$\text{frac}_{\text{spatial}} = S_{\upsilon}{}^{2} \big/ (S_{\upsilon}{}^{2} + \sigma_{v}{}^{2})$$

$$\overline{\upsilon} = \text{average of } \upsilon$$

$$\sigma_{v}{}^{2} = \text{variance of marginal unstructured component}$$

$$\sigma_{\upsilon}{}^{2} = \text{variance of marginal structured component}$$

To evaluate predictive performance, district specific predicted posterior mean values from both the state-level and the South India model were compared with the corresponding observed mean number of outbreaks using pair-wise Pearson's correlation statistics. To test the out-of-fit predictive performance of the model, leave-one-out cross validation statistics, namely Conditional Predictive Ordinates (CPOs) were calculated (Gelfand, 1996). The CPO expresses the posterior probability of observing the value of y_i when the model is fitted to all data except y_i, with a larger value implying a better fit of the model to y_i, and very low CPO values suggesting that y_i is an outlier and an influential observation resulting in bias in estimates. When many CPO values cluster near zero, the model demonstrates poor out-of-fit performance. When many CPO values cluster near one, the model demonstrates good out-of-fit performance (Lawson, 2013).

4.3 Results

The mean annual number of outbreaks per district ranged from 0 to 33.83 (mean ± s.e. = 3.74 ± 7.13). The district with the most annual outbreaks was Prakasham (Andhra Pradesh) with 370 outbreaks in 1998. The mean annual number of outbreaks per district ranged from 0 to 19 (mean ± s.e. = 2.11± 4.59) in Karnataka from 0 to 28.05 (mean ± s.e. = 3.27±7.25) in Tamil Nadu and from 0 to 33.83 (mean ± s.e. = 3.74 ± 7.13) in Andhra Pradesh.

4.3.1 South India model

Models containing only land cover or host predictors out-performed models with only climate predictors (Table 4.3) and models with spatial random effects outperformed models with non-spatial random effects in all geographical areas.

Considering the best combined models, the number of outbreaks in South India (Table 4.5), were positively associated with rain fed croplands and the Nellore breed of sheep, and negatively associated with closed to open (>15%) shrub land and South-West monsoon rainfall. The structured and unstructured spatial components were significant in the combined best model (Table 4.5).

Comparison of different models with different structures within individual categories of models (host and land cover), resulted in better performance of models with BYM and Besag structure than with *iid* models. Within the climate models, however, the BYM and *iid* models outperformed the Besag model. The model with covariates only was the worst performing model in all the individual categories of models. The BYM model had a lower DIC (DIC=199.27) than models containing either the predictors alone (DIC=486.46) or spatial random effects alone (DIC= 231.25) (Table 4.4), and had fewer outliers (detected from CPO and cross-validation statistics). A large proportion of the variance described by the spatial random effects was explained by unstructured heterogeneity (>99%) rather than structured heterogeneity.

The predicted mean number of outbreaks using the best model (Fig. 4.4B) shows excellent correspondence with the observed mean number of outbreaks (Fig. 4.4A), with a correlation co-efficient of $r = 0.996$ ($p < 0.005$) for a BYM model and a correlation co-efficient of $r = 0.68$ ($p<0.005$) for a covariates only model (without random effects). The model is able to discriminate between districts with and without outbreaks and also delineates severely affected districts successfully. Areas with a reported absence of BT include regions with forest cover (Fig. 4.2B) at high altitude which have few sheep.

4.3.2. State-level models

The host and land cover models within individual categories outperformed the climate model in all the three states, but host variables outperformed land cover variables and climate variables in the Karnataka models. Considering the best combined models (Table 4.6) for Andhra Pradesh, the mean annual numbers of BT outbreaks were significantly and positively related to the abundance of buffalo, exotic & crossbred sheep and goat populations and to artificial surfaces and associated areas, and significantly negatively related to the area of closed broadleaved deciduous forest. The average numbers of BTV outbreaks in Karnataka (Table 4.7) were significantly positively related to abundance of non-descript sheep and exotic & crossbred sheep and significantly negatively related to the abundance of Mandya sheep. In Tamil Nadu BTV outbreaks were significantly positively associated with the abundance of Ramnad white sheep and significantly negatively associated with the North-East monsoon rainfall (Table 4.8).

Comparison of different models with different structures within individual categories of models resulted in a slightly better performance of the BYM and *iid* models than the Besag model in Andhra Pradesh and Karnataka. Only in Tamil Nadu did the BYM and Besag model outperform the *iid* model. Within the climate models, the Besag model outperformed the BYM and *iid* models in Andhra Pradesh and Tamil Nadu. The model with covariates only was the worst performing model in all the individual category models (Table 4.7) in all the three states.

The predicted mean number of outbreaks using the best BYM state level models including covariates show excellent correspondence with the observed mean number of outbreaks for all three states, with a correlation co-efficient of r =0.983 for Andhra Pradesh, r = 0.991 for Karnataka and r = 0.997 for Tamil Nadu.. The models with covariates only (i.e. no random effects) show better correspondence with observed outbreaks for Andhra Pradesh and Karnataka (correlation co-efficient of r = 0.76 for Andhra Pradesh and r = 0.80 for Karnataka than for Tamil Nadu (correlation co-efficient of r = 0.60).

Land cover variables	South India	AP	Karnataka	Tamil Nadu
Post-flooding or irrigated croplands	X	X	X	X
Rain fed croplands	X	X	X	X
Mosaic cropland (50-70%)/vegetation (grassland, shrubland, forest (20-50%)	X	X	X	X
Closed to open (>15%) broadleaved evergreen and/or semidecidous forest (>5m)	X	X	X	X
Closed (>40%) broadleaved deciduous forest (>5m)	X	X	X	X
Closed (>40%) needle-leaved evergreen forest (>5m)	X	X	X	X
Closed to open(>15%) shrub land(<5m)	X	X	X	X
Closed to open (>15%) grassland	X	X	X	X
Artificial surfaces and associated areas (urban areas>50%)	X	X	X	X
Water bodies	X	X	X	X

Table 4.1: Land cover variables considered in the combined South India model and individual state level models. X indicates variable was considered in the variable selection.

Host and breed Variables	South India	Andhra Pradesh	Karnataka	Tamil Nadu
Goat	X	X	X	X
Sheep breeds				
Exotic and cross bred	X	X	X	X
Bellari	X	-	X	-
Coimbatore	X	-	-	X
Deccani	X	X	X	-
Hassan	X	-	X	-
Kengur	X	-	X	-
Kilakarsal	X	-	-	X
Madras Red	X	-	-	X
Mandy breed	X	-	X	-
Mercheri breed	X	-	-	X
Nellore breed	X	X	-	-
Nilgiri breed	X	-	-	X
Ramnad white breed	X	-	-	X
Non-descript	X	X	X	X
Tiruchi.black	X	-	-	X
Vembur	X	-	-	X
Indigenous cattle	X	X	X	X
Cross-bred cattle	X	X	X	X
Buffalo	X	X	X	X

Table 4.2: Host and breed variables considered in the combined South India model and individual state level models. X indicates variable was considered in the variable selection; - indicates there were few/none of these particular hosts in each region and that these hosts were therefore not considered in the variable selection.

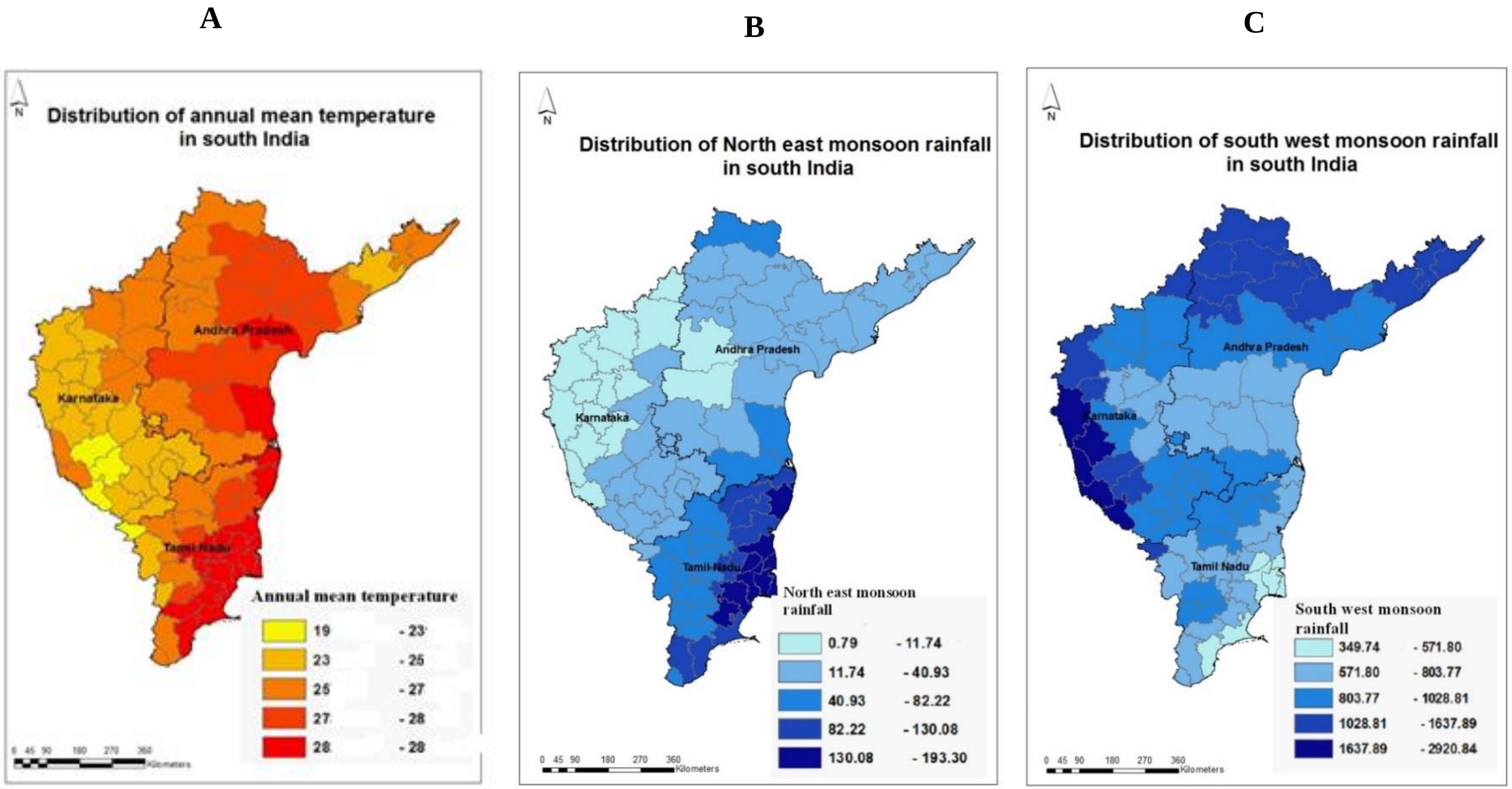


Figure 4.1: Maps of climate variables in southern India: (A) Annual mean temperature (^{0}C) (B) Annual average North-East monsoon precipitation (mm) and (C) Average South West monsoon precipitation (mm) (obtained from Xie et al., 2010).

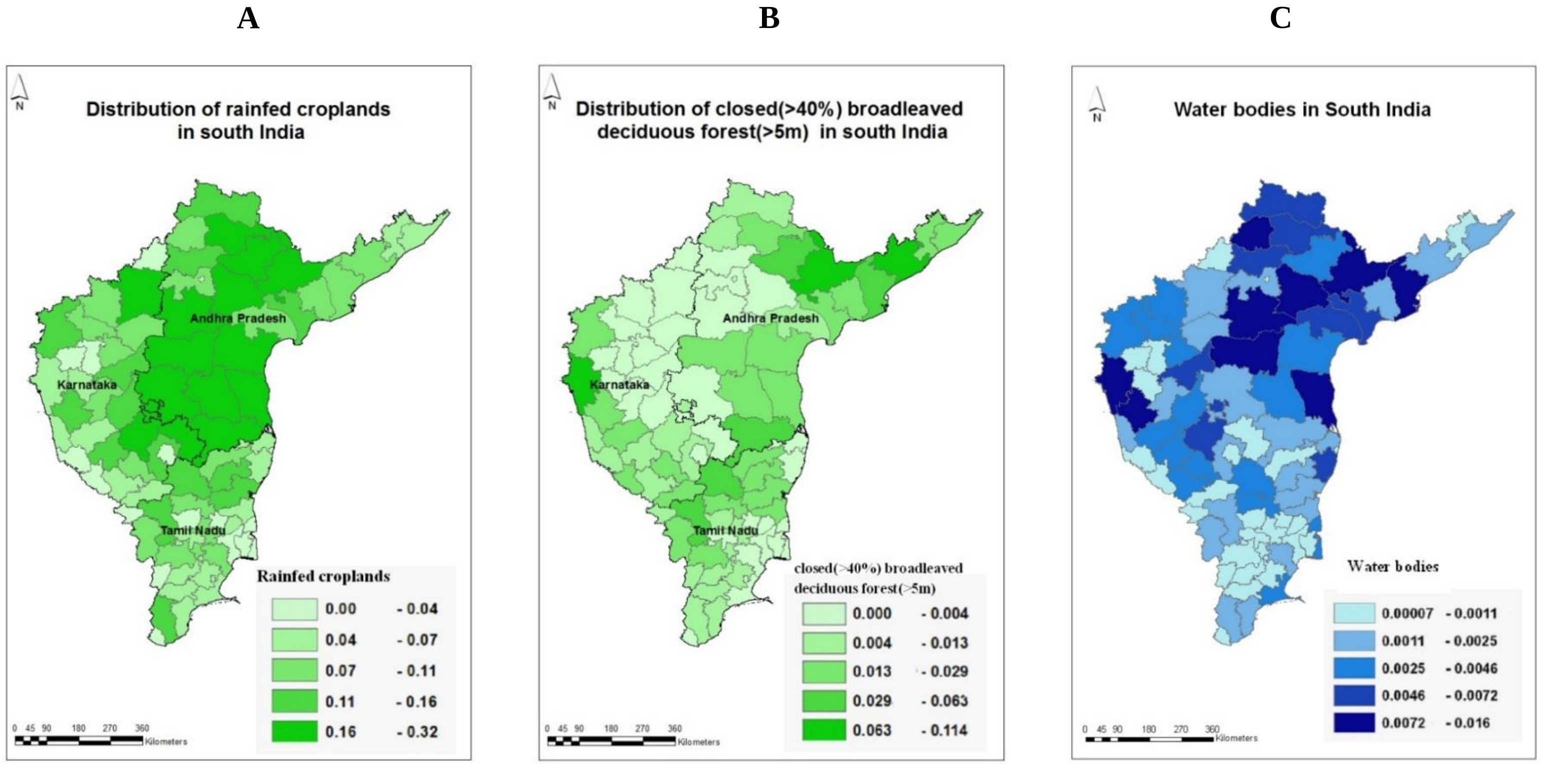


Figure. 4.2 Maps of district level land use variables in southern India: (A) Rain fed croplands; (B) closed (>40%) broadleaved deciduous forest (>5m) and (C) Water bodies (all obtained from Defourny et al., 2006))

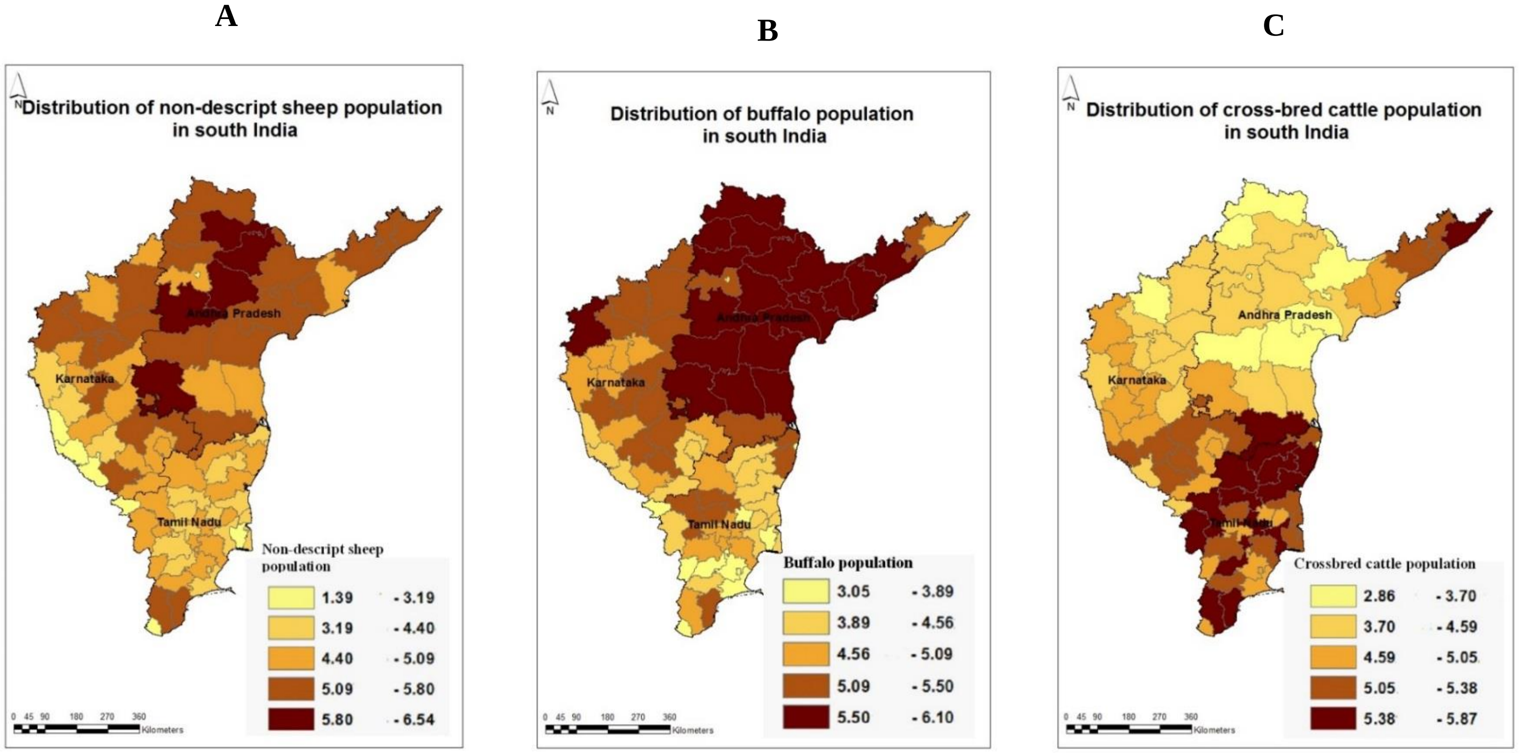

A
B
C
Distribution of non-descript sheep population in south India
Andhra Pradesh
Karnataka
Tamil Nadu
Non-descript sheep population
1.39 - 3.19
3.19 - 4.40
4.40 - 5.09
5.09 - 5.80
5.80 - 6.54
0 45 90 180 270 360 Kilometers
Distribution of buffalo population in south India
Andhra Pradesh
Karnataka
Tamil Nadu
Buffalo population
3.05 - 3.89
3.89 - 4.56
4.56 - 5.09
5.09 - 5.50
5.50 - 6.10
0 45 90 180 270 360 Kilometers
Distribution of cross-bred cattle population in south India
Andhra Pradesh
Karnataka
Tamil Nadu
Crossbred cattle population
2.86 - 3.70
3.70 - 4.59
4.59 - 5.05
5.05 - 5.38
5.38 - 5.87
0 45 90 180 270 360 Kilometers

Figure. 4.3 Maps of district level host abundance in southern India (all on a log$_e$ scale): (A) non-descript sheep (log.$_e$); (B) buffalo (log.$_e$); and (C) crossbred cattle (log.$_e$) (obtained from National livestock census 2007).

Model	Predictors in model	DIC (BYM+covariate)	DIC (besag +covariate)	DIC (iid +covariate)	DIC (covariate only)	BYM only
Host	Crossbred cattle, Buffalo, non-descript sheep, Bellary breed of sheep, Nellore breed of sheep	202.96	202.57	207.44	600.99	231.25
Land-cover	Post-flooding or irrigated croplands, Rain fed croplands, closed to open (>15%) shrub land (<5m), artificial surfaces and associated areas (urban areas >50%), water bodies	199.74	199.40	204.62	584.76	
Climate	North East monsoon, annual monsoon, annual mean temperature	216.27	219.08	216.00	670.87	

Table 4.3: Deviance information criterion (DIC) for the best models of bluetongue severity across South India, where predictors are drawn from a single category of environmental predictors. Comparison between Besag-York-Mollie (BYM) model with covariates, Besag model, iid model, covariate only model and BYM model without covariates.

Model No.	Model	DIC	pD
1.	BYM + South West monsoon rainfall, rain fed croplands, closed to open (>15%) shrub land (<5m), water bodies, Bellari breed of sheep, non-descript sheep, buffalo, crossbred cattle and Nellore	196.84	43.67
2.	South West monsoon rainfall, rain fed croplands, closed to open (>15%) shrub land (<5m), water bodies, Bellari breed of sheep, non-descript sheep, buffalo, crossbred cattle and Nellore	486.46	9.963
3.	BYM only model (no covariates)	231.25	61.48

Table 4.4: Final model of bluetongue severity across south India which includes the best combination of host, climate and landscape predictors (top row). The DIC and pD (Effective number of parameters) for this model (BYM + covariates, Model 1) are compared to models containing BYM only (Model 3) or covariates only model (Model 2).

Effects	Mean(sd)	Credible interval
Fixed effects		
Intercept	-27.32(11.63)	-51.72, -5.87
South West monsoon rainfall	-0.0017 (0.0008)	**-0.0035, -0.0002**
Rain fed croplands	1.54 (0.51)	**0.54, 2.59**
Closed to open (>15%) shrub land(<5m)	-2.59(1.25)	**-5.17, -0.20**
Water bodies	-4.14 (2.69)	-9.74, 0.87
Bellary breed of sheep	0.15 (0.13)	-0.09, 0.41
Non-descript sheep	0.26 (0.45)	-0.60, 1.19
Buffalo	0.64 (0.46)	-0.24, 1.58
Crossbred cattle	0.59 (0.40)	-0.18, 1.43
Nellore breed of sheep	0.37 (0.14)	**0.094 , 0.67**
Random effects		
Spatial component	0.26 (0.08)	**0.13 , 0.48**
Unstructured component	1898.39(1854.17)	**125.70, 6760.08**

Table 4.5: Mean coefficient values and credible intervals for fixed effects environmental predictors which describe the average annual number of bluetongue outbreaks per district for the best model for South India. Significant credible intervals (i.e. those that do not span the value zero) are in bold.

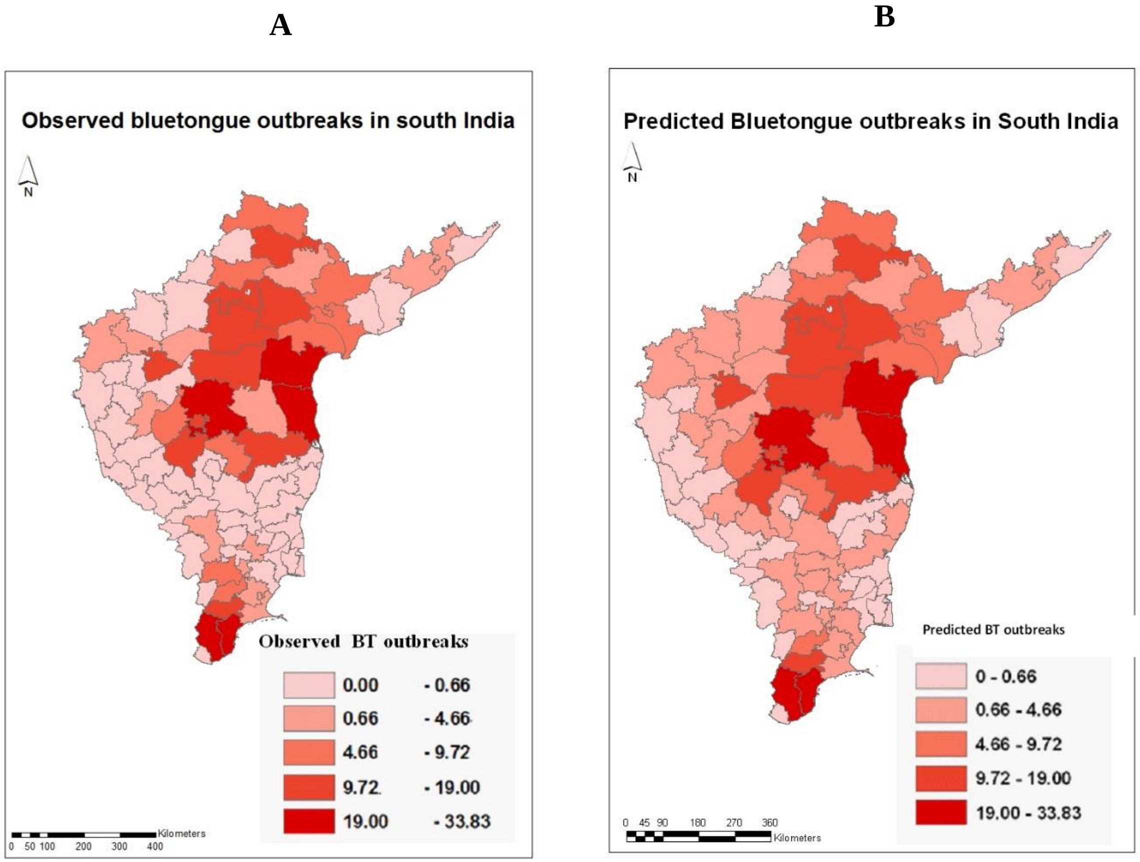


Figure. 4.4 (A) Observed average annual number of outbreaks in South India from PD_ADMAS; and (B) Predicted average annual number of outbreaks from the best model.

	Mean(sd)	**Credible interval**
Fixed effects		
Intercept	-8.37(6.57)	-22.31, 3.86
Buffalo	1.51(0.57)	**0.43, 2.72**
Exotic& crossbred sheep	0.3091(0.08)	**0.14, 0.49**
Goat	2.13(0.76)	**0.76, 3.81**
Closed (>40%) broadleaved deciduous forest (>5m)	-0.98(0.33)	**-1.68, -0.34**
Artificial surfaces and associated areas (urban areas>50%)	3.93(1.68)	**0.68, 7.36**
Random effects		
Precision for unstructured component	1805.29(1780.88)	**116.35, 6531.88**
Precision for structured component	2.16 (1.59)	**0.49, 6.36**

Table 4.6: Mean coefficient values and credible intervals for fixed effects environmental predictors which describe average annual number of bluetongue outbreaks in Andhra Pradesh for the best subset models across the land-cover, climate and host categories. Significant credible intervals (i.e. those that do not span the value zero) are in bold.

	Mean(sd)	**Credible interval**
Fixed effects		
Intercept	-21.80(7.30)	**-38.14, -9.24**
Non-descript sheep	3.00(0.94)	**1.30, 5.06**
Exotic and crossbred sheep	1.16(0.35)	**0.52, 1.91**
Mandya breed of sheep	-0.75(0.31)	**-1.45, -0.18**
Crossbred cattle	1.25(0.99)	-0.47, 3.48
Random effects		
Precision for unstructured component	1.43 (1.019)	**0.33 , 4.10**
Precision for structured component	1858.04 (1836.24)	**126.72, 6698.38**

Table 4.7: Mean coefficient values and credible intervals for fixed effects environmental predictors which describe average annual number of bluetongue outbreaks in Karnataka for the best subset models across the land-cover, climate and host categories. Significant credible intervals (i.e. those that do not span the value zero) are in bold.

Fixed effects	**Mean(sd)**	**Credible interval**
Intercept	-15.53(9.17)	-35.38, 0.92
Nilgiri breed of sheep	0.42(0.26)	-0.07, 0.97
Ramnad white breed of sheep	0.85(0.27)	**0.35, 1.43**
North –East monsoon rainfall	-0.03(0.01)	**-0.05, -0.01**
Annual mean temperature	0.05(0.03)	-0.008, 0.12
Precision for unstructured component	1819.30(1774.23)	**122.71, 6530.29**
Precision for structured component	0.54 (0.28)	**0.16 , 1.26**

Table 4.8: Mean coefficient values and credible intervals for fixed effects environmental predictors which describe average annual number of bluetongue outbreaks in Tamil Nadu for the best subset models within the land-cover, climate and host categories. Significant credible intervals (i.e. those that do not span the value zero) are in bold.

4.4 Discussion and Conclusion

This study is the first of its kind to explain spatial patterns in BT severity in South India in relation to a full range of important environmental variables. The results indicate that host and landscape heterogeneity are much more important in determining spatial patterns in BT severity in South India than is spatial heterogeneity in climate conditions. Although monsoon conditions undoubtedly contribute to the disparity in severity of BT between North India (little affected by monsoons) and South India (heavily affected by monsoons, see Prasad et al., 2009), the analysis here indicates that host and landscape predictors come into play at finer spatial scales. The finding that models combining different suites of environmental predictors (namely host, climate and landscape predictors) out performed those based on single suites of predictors illustrates the value of considering all potential environmental variables in the same model framework (Acevedo et al., 2010; Purse et al., 2012).

Considering the combined South India model, districts with a higher land-cover of rain fed croplands suffered significantly more BT outbreaks, probably because such landscapes are more likely to contain suitable breeding habitats and hosts of the *Culicoides* biting midge vectors. Although the vectorial capacity of the different *Culicoides* species has not been well studied in India, three key species, *C. imicola*, *C. peregrinus*, and *C. oxystoma* seem to be abundant in BT-affected areas (Reddy & Hafeez, 2008). Populations of *C. imicola* and *C. peregrinus* that both breed in moist soil are significantly associated with irrigated areas or areas with high soil moisture availability elsewhere (Acevedo et al., 2010). *C. oxystoma*, which breeds in buffalo dung (Narladkar et al., 2006), and its larvae, have been found in both active and abandoned rice paddy fields (encompassed by the irrigated cropland class) elsewhere in Asia (Yanase et al., 2013). Surprisingly, areas with relatively high annual monsoon rainfall in South India had lower numbers of outbreaks. This may be because these higher rainfall regions (Fig. 4.1c), for example the western Ghat forests, support very low densities of sheep and other livestock, and therefore probably support low populations of livestock-associated *Culicoides* species, leading to fewer or no BTV outbreaks.

The finding of different variables selected in each state supports our rationale behind fitting state level models. In the Andhra Pradesh model, host and land cover variables dominate and climate variables are less important. The host and breed variables dominate in the Karnataka model over land cover or climate variable models. Finally, in the Tamil Nadu model, host and climate variables were more important than land cover variables. Except for exotic & crossbred sheep selected in Andhra Pradesh and Karnataka, other host variables were not shared in common with the other state models. Land cover variables were selected only in Andhra Pradesh and not in Karnataka and Tamil Nadu. Climate variables were selected only in Tamil Nadu (North-East monsoon rainfall and Annual mean temperature) and not in the other two states.

Buffalo and goat are more abundant in Andhra Pradesh than in Karnataka or Tamil Nadu. The Mandya breed of sheep (selected in Karnataka) is absent in Andhra Pradesh and Tamil Nadu. Similarly, the Nilgiri breed and Ramnad white breed of sheep (present in Tamil Nadu) are absent from the other two states. Tamil Nadu receives the highest North East monsoon rainfall compared to the other two states (Fig 4.1B).

In Andhra Pradesh, BT outbreaks at the district level were negatively associated with higher percentage coverages of closed broadleaved deciduous forest (Fig. 4.2B). This again may be because such habitats are less suitable for *Culicoides*, or contain fewer hosts.

The positive effects of exotic and crossbred sheep on outbreak numbers in Karnataka and Andhra Pradesh are consistent with previous findings of the high susceptibility of such breeds in India (Lonkar et al., 1983), and the restriction of past disease cases in South-East Asia to European sheep breeds (Daniels et al., 2003). Antibodies against BTV are reported in local breeds in Indonesia and Malaysia without any clinical signs of disease (Hassan et al., 1992; Sendow et al., 1991), suggesting that local breeds are susceptible to infection (and hence may act as efficient hosts of BTV), with the result that BT presence, when recorded only on the basis of clinical signs (the most frequent way of reporting BT across the whole of India), will be under-reported where these breeds are common. The strong and positive association of BTV outbreaks with goat populations in Andhra Pradesh is interesting because small and marginal farmers practice mixed farming of sheep and goat with other livestock. A high sero-prevalence in this species (goat) without clinical signs has been reported by (Arun et al., 2014; Bitew et al., 2013). A similar finding applies to buffaloes (positively associated in the models with BTV outbreaks in AP only), which also demonstrate a high seroprevalence without clinical disease (Kakker et al., 2002). Thus, both goats and buffaloes may be important reservoir hosts of BTV in Andhra Pradesh. It is probable that the extensive rice belt found in Andhra Pradesh (the state most severely affected by BT), that supports high buffalo populations (Fig. 4.4 A) and likely high populations of *C.oxystoma*, makes a substantial contribution to maintaining BT transmission.

Different local breeds in each state seem to be important in the local transmission of BTV. Nellore sheep abundance was a significant variable in the Andhra Pradesh model (Table 4.4). A more clinically severe form of the disease has been reported in this breed compared to other breeds (Rao et al., 2014).

The selection of the Ramnad white breed of sheep in Tamil Nadu is interesting because this breed along with Trichy black has also been shown to be more susceptible to bluetongue compared to other indigenous breeds (Prasad et al 2009; Rao et al., 2014).

A large proportion of the variance described by the spatial random effects was explained by structured heterogeneity (>99%, versus the unstructured heterogeneity) in the South India model. This suggests that either intrinsic processes such as disease spread between districts, or unmeasured spatially structured environmental predictors drive district-level disease patterns. The latter could include soil (soil type and water retention capacity) or animal husbandry factors (dung management and use as fertiliser, local drainage and flooding), that may influence breeding site availability and abundance of potential vectors. The importance of spatial structure is also supported by the better performance of the BYM and Besag models over the *i.i.d.* models (Table 4.1). The contribution of different random effect components (structure and unstructured heterogeneity) also varies with each state. Structured heterogeneity is important in the Andhra Pradesh and Tamil Nadu models compared to unstructured heterogeneity within each state, and vice-versa in Karnataka. The probable reasons for dominance of unstructured heterogeneity in Karnataka is the influence of different farm/village level practices which influence the breeding of midges and also their abundance.

Thus, state-specific models are required to understand the role of different variables in determining the severity of BTV outbreaks in South India and also other states of India which are not only different in their ecoepidemiological factors but also different in their disease surveillance systems, the latter partly because of the varying number of veterinary hospitals and disease diagnostic laboratories in each state. The use of state-specific models is also justified because of the existence of different reporting systems and awareness of bluetongue virus in the different states.

Chapter 5

Understanding the inter-annual spatio-temporal risk factors for bluetongue outbreaks across South India using Bayesian Poisson regression modelling.

5.1 Introduction

The role of climate in driving inter-annual variability of bluetongue outbreaks in space and time is not well known globally, especially in endemic countries (Coetzee et al., 2012). A sequence of droughts followed by floods seems to explain the variability of a related midge-borne orbivirus, African horse sickness virus (AHSV), in Africa (Baylis et al., 1999b). The factors determining the inter-annual variability in the BTV outbreaks in South India at district level are not well known, but may include long term changes in climate, particular sequences of dry and wet years, or a waxing and waning of herd immunity. The role of climate along with host and land cover in determining spatio-temporal variability in outbreaks is addressed in this Chapter by using a Bayesian Generalised Linear Mixed model accounting for spatial and temporal autocorrelation. The resulting model will be helpful in making forecasts and possibly in the development of an early warning system for the disease in India.

Understanding the spatial and temporal epidemiology of BTV in South India independently has helped to identify the mechanisms and factors responsible for variation between districts in space and over time within a state (Chapters 4 and 3 respectively). Seasonality of outbreaks is driven by precipitation, whereas spatial heterogeneity is driven by a combination of host, land cover and climate factors. Furthermore, the previous Chapters have shown how BTV outbreaks could be adequately forecast in space and time independently. What is important to establish now is whether or not, and if so how, these factors operate together in space and time. This will help us to predict BTV outbreaks more accurately in the future.

Climate heterogeneity, especially monsoon variability in space and time in India is well known (Kumar et al., 1992). There are many studies showing the importance of climate in vector-borne disease outbreaks, an importance which is made very clear by the basic Reproductive number (R_0) formula for such diseases (Hartemink et al., 2009), in which many parameters and variables refer to demographic and other processes in the life cycle of the vectors which are very susceptible to climate. The quantification of the relationships between weather, climate and vector-borne diseases naturally leads to the development of Disease Early Warning Systems (DEWS) to forecast such diseases in operationally useful ways. Early Warning Systems (EWS) in general have been developed for forecasting famines, forest fires and hurricanes (Roger, 1997). Forecasting infectious disease outbreaks goes back to as early as 1920's in India, where malaria was predicted on a district by district basis using data on past malaria outbreaks, market prices of food and long term meteorological data (Myers et al., 2000). This application - probably the best example of a long term, accurate space and time disease forecasting service – fell into disuse at the end of the 1940s when alternative methods (e.g. new insecticides) became available for combating the disease. Relatively few DEWS have been developed since the 1940s and the majority of these make predictions in the time domain only (Chaves & Pascual, 2007; Medina et al., 2008) and more rarely in both the space and time domains (Thomson & Palmer et al., 2006).

The role of El-Nino in outbreaks of vector borne diseases has been reported in many studies (Hales et al., 1999; Chaves & Pascual 2007), but it is unlikely that El-Nino itself has a direct impact on the diseases concerned, instead acting indirectly via its effects on important climate variables, especially rainfall, floods and droughts (Dilley & Heyman, 1995). In other work, a strong and significant association was found between a particular sequence of drought and flooding events in 13 of the 14 African Horse sickness virus outbreaks over two centuries in Africa. This sequence of events was attributed to the El Niño/Southern Oscillation (ENSO) (Baylis et al., 1999b). There is a report of strengthening of relationship between ENSO and North-East monsoon rainfall in South India and Sri Lanka (Zubair & Ropelewski, 2006) compared to the weakening of the relationship with South-West monsoon (Kumar et al., 1999) and there was significant correlation between Nino-3 ENSO index and North-East monsoon rainfall in comparison to the past ENSO events (1982, 1987, 1997) (Zubair & Ropelewski, 2006). El-Nino's role in the Indian monsoon, especially in South India, has already been demonstrated (Annamalai et al., 2007; Rasmusson & Carpenter, 1983). The relationship between ENSO and the All India Rainfall index (AIR), which is area-weighted seasonal average of South-West monsoon rainfall is predominant. The AIR index is typically (not always) below (above) normal during El Niño (La Nina) years (Annamalai et al., 2005).

It is important to understand the purpose, scale and feasibility of implementing an operational DEWS. There needs to be systematic integration of epidemiological surveillance data, environmental and other observations and they should be assessed within a model framework (Burke et al., 2001). Currently, a system exists in India for forecasting the presence or absence of bluetongue and other diseases two months in advance (www.nadres.in) at the district level. The system does not predict the number of outbreaks of bluetongue and the model performance statistics are not documented.

Host and land cover heterogeneity was important in determining the spatial risk of bluetongue (Chapter 4), but whether or not these factors also play an important role when considered as static variables in spatio-temporal models is not known. Thus, quantifying the role of climate, host and land cover in determining the inter-annual variability of bluetongue outbreaks in each district in different years will help to develop early warning system for the disease in India.

5.1.1 Space-time analysis of epidemiological data

Regression methods for spatio-temporal analysis depend on whether the response variable is count data, or presence and absence data, or a continuous variable. Count data are most often modelled assuming a Poisson distribution in a standard GLM (Generalised Linear Model). The Poisson distribution assumes that the mean and variance are equal; when they are not equal there is either under dispersion (sample variance<mean) or over dispersion (sample variance>mean). Aggregated count data in epidemiology are often over dispersed (Clements et al., 2006) and this over dispersion is sometimes referred to as extra-Poisson variability. Failure to account for over dispersion can lead to bias in estimates. The extra-Poisson variability can arise for many reasons, for example due to the presence of spatial and temporal autocorrelation in the situation where the data were gathered (country, district or region; over days, months or years) or to the lack of information about an important predictor variable in space and time (unobserved covariate).

Extra-Poisson variability can be accounted for by using modifications of the Poisson model such as zero-inflated models (Zuur et al., 2009) or by choosing an alternative model such as a negative binomial one (Clements et al., 2006). Both sorts of model are capable of generating over dispersion, where there are more zeroes (no outbreaks) and more very high values (multiple outbreaks) than is the case for a Poisson distribution with the same mean. The negative binomial distribution model is a modification of the Poisson model involving an extra parameter (commonly given the symbol 'k') to account for the obvious 'clumping' of the data (k=0 for highly clumped data and k =∞ for the Poisson distribution). Interpretation of extra-Poisson variability is often difficult in zero-inflated or negative binomial model outputs. In addition, the same results can be obtained by combining a series of Poisson distributions each with a different mean (and hence variance).

Extra-Poisson variability can be accounted for by including spatial and temporal autocorrelation in a Generalised linear mixed model (GLMM). The spatial and temporal autocorrelation parameters can be modelled as fixed parameters using maximum likelihood estimation in the frequentist domain or as random parameters in the Bayesian domain (Clayton, 1996).

The presence of extra-Poisson variability, even after accounting for spatio-temporal autocorrelation, may lead to bias in estimates. Extra-Poisson variability can arise due to the absence of one or more predictor variables which interact in both spatial and temporal domains (structured interaction) or it can arise due to absence of one or more predictor variables which interact independently, without spatial and temporal structure. These interaction terms cannot be modelled using maximum likelihood approaches, but can be modelled in a Bayesian hierarchical framework by specifying priors for each parameter (spatio-temporal interactions).

There are many ways of introducing space-time interactions in Bayesian disease mapping (Assunção et al., 2001; Martínez-Beneito et al., 2008). Bayesian Generalized linear mixed models (GLMM) using the Besag-York-Mollie (BYM) (Besag & Newell, 1991) model, discussed in chapter 4, can be extended in space-time by incorporating temporal dependence (Bernardinelli et al., 1995; Richardson et al., 2006). The basic BYM space-time model assumes that spatial and temporal dependence act independently of each other. An extension of the model overcomes this problem by separately modelling spatial dependency, temporal dependency and different interaction terms through prior distributions (Knorr-Held, 1999).

There are many advantages of Bayesian space-time methods to account for spatial and temporal heterogeneity in species' distribution modelling (Gelfand et al., 2005), disease ecology (Waller et al., 2007) and wildlife diseases (Farnsworth et al., 2006). In studies of disease ecology, such as those for Lyme disease (Waller et al., 2007), the inclusion of spatio-temporal components in a Bayesian framework resulted in better understanding of disease spread in space and time than in the model without spatio-temporal structure.

The Bayesian GLMM (with interaction terms) can account for different heterogeneities and missing covariates as well as the fixed effects of environmental predictors and was therefore used in the present analysis. This chapter uses the Bayesian GLMM model fitted in INLA (Integrated Nested Laplace Approximation) (Blangiardo et al., 2013), to investigate the role of climate, land cover and host variables in the inter-annual variability of BTV outbreaks in South India, in order to answer the following questions

1. Are outbreaks restricted in either time or space (i.e. to certain times, or to certain districts only), or do they occur randomly in both time and space?
2. Is the severity of BTV outbreaks greater in years when there is a sequence of extreme events such as dry years (high temperature and low rainfall) followed by wet years (low temperature and high rainfall)?
3. Is there any significant correlation between periodicities in the Sea Surface Temperature (SST, one of the measured El Nino variables) and different monsoon conditions and, if so, how does it vary between the three states of South India?
4. Do the roles of host and land cover remain strong (as in spatial analysis) in describing inter-annual variability in bluetongue outbreaks in different districts when considered as static variables?

5. Can adequate forecasts (with low RMSE error and high correlation between observed and predicted outbreaks on out-of-fit data) be made for use in early warning systems for the disease in India?

5.2 Materials and Methods

5.2.1 Disease data

District level (admin-2) monthly BT outbreak data (1992-2009) were provided by PD_ADMAS (Project Directorate on Animal Disease Monitoring and Surveillance) which maintains the livestock diseases database for India and collates outbreak data every month from different sources. The monthly data were aggregated to a yearly level and the sum of all bluetongue outbreaks occurring each year in each district was modelled as the dependent variable. The district level outbreak data were summarised at the annual level because there was sparseness in the outbreaks (there were no outbreaks in many districts and months). These yearly bluetongue outbreak data were divided into training (1992-2007) and test data (2008 and 2009) to test model accuracy. In the time series analysis (chapter 3), the years 2008 and 2009 were not included in the 'out-of-fit' forecast because only Andhra Pradesh (no outbreaks reported in 2009) was included in the analysis, but in this chapter the other two states (Karnataka and Tamil Nadu) are included in the analysis and there were reports of BTV outbreaks in these two states for the years 2008 and 2009.

5.2.2 Predictor variables

Yearly annual mean maximum temperature (i.e. the average of the 12 monthly maximum temperatures) and the yearly sum of monthly precipitation data (1992-2009) were obtained from the Climatic Research

Unit (CRU) (5°×5° gridded data), University of East Anglia, UK (Harris et al., 2014) and extracted in the zonal statistics option in ArcMap 10.1 (ESRI, Inc., Redlands, CA, U.S.A.) by specifying districts (admin level 2). The annual mean maximum temperature (^{0}C) and annual total rainfall (mm) were each lagged by zero, one and two years.

State-wide annual monsoon rainfall (Jan-Dec), South-West monsoon rainfall (June-Sept) and North-East monsoon data (Oct-Dec) from 1901-2000 were purchased from the Indian Meteorological Department for investigating the correlation between dominant frequencies of the rainfall and Sea Surface Temperature (SST) data. Rainfall data were used for the years 1949-2000 because the El-Nino data are only available from 1949. Hence the wavelet analysis was restricted to these years.

The term ‘El-Nino’ is applied to the warm phase of ENSO and ‘La Nina’ to the cold phase of ENSO. The Nino 3 region covers the geographical region between 5^{0}N-5^{0}S, 90^0-150^{0}W, i.e. a ten degree equatorial band in the eastern Pacific Ocean. The monthly sea-surface temperature (SST) data (1949-2000) were obtained from the Japan Meteorological Agency (JMA) (Ishii et al., 2005) as mentioned in chapter 3. The annual averages of the monthly sea surface temperature (El-Nino 3) data were calculated and used in the analysis.

The proportional area of each district covered by ten land-cover classes as described in the spatial analysis Chapter were extracted from the GlobCover land-cover map (Defourny et al., 2006) using zonal Statistics in ArcMap 10.1 (ESRI, Inc., Redlands, CA,U.S.A.), and were logit-transformed. These ten classes were considered for selection in the individual land-cover category model.

The host variables used in this analysis (in each case total numbers) were; local breeds of sheep, exotic & cross bred sheep, non-descript sheep, indigenous cattle, crossbred cattle, buffalo and goat population. The indigenous breeds encompass all the fourteen local breeds of South India. The breeds selected in the spatial analysis were specific to each state, therefore it was decided to combine all the local breeds from three states of South India to be considered in the variable selection procedure. The host variables were used as static variables for all the years under study, because the livestock census is carried out only every five years.

5.2.2 Exploratory data analysis

Displaying data in space and time is important to discriminate between endemic and hyper endemic years. It also helps to understand the changes over time and also the spatio-temporal patterns in the data. The space-time multi panel plots were generated using the **space-time** package in R (Pebesma, 2012). The yearly mean of the outbreaks in all the districts were calculated to determine the criteria for deciding whether a year was hyper-endemic or endemic. Years with means above the overall mean were considered as hyper-endemic years and years with means equal to or below the overall mean were considered as endemic years.

Wavelet analyses and wavelet coherence (correlation) analyses of annual average El-Nino 3 temperature data and annual & seasonal (South-West monsoon, North-East monsoon) rainfall data were conducted for the three different states separately to identify the correlation between the local dominant frequencies. The annual and seasonal rainfall variables for Andhra Pradesh, Karnataka and Tamil Nadu were subjected to wavelet coherence analysis with the sea surface temperature data to identify any significant wavelet coherence (correlation) and whether the correlation changes with each state.

5.2.3 Modelling approach

Relationships between the annual number of outbreaks in each district and environmental factors were quantified using a Generalised linear mixed model with Poisson errors, implemented in a Bayesian framework. Different models were fitted, with different spatial and temporal random effects and interactions of these effects. The probability function for Y, the mean number of BTV outbreaks, is given by

$$Y_{it} \mid \mu_{it} \sim Poisson(E_{it}\mu_{it})$$

$$\log \mu_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_n x_{nit} + \varepsilon_i + s_i + \gamma_t + s_t$$

$$\left.\begin{aligned} &\varepsilon_i \mid \varepsilon_j = 1 \sim Normal(m_{i,} s^2{}_i) \\ &m_i = \frac{\sum_{j \in N(i)} \varepsilon_j}{\neq N} \\ &s_i^2 = \frac{\sigma_\varepsilon{}^2}{\neq N_{(i)}} \end{aligned}\right\} \text{Spatial structured and unstructured heterogeneity} \tag{1}$$

$$\left.\begin{aligned} &\gamma_t = \sum_{j=1}^{p} \rho_j \gamma_{t-1} + Z_t \\ &Z_t \sim N(0, \sigma^2 \gamma) \\ &\tau_t \mid \tau_j = 1 \sim Normal(m_t, s^2{}_t) \\ &m_t = \frac{\sum_{j \in N(t)} \tau_j}{\neq N} \\ &s^2{}_t = \frac{\sigma_\tau{}^2}{\neq N_{(t)}} \end{aligned}\right\} \text{Temporal structured and unstructured heterogeneity}$$

Where β_0 is the intercept and β_1 to β_n are the co-efficients for the fixed effects of predictors (x_{1it} to x_{nit}) every year in each district. ε_i is the structured spatial component assuming Besag-York-Mollie (BYM) specification (Besag et al 1991), modelled using an intrinsic autoregressive structure (iCAR). $\neq N_{(i)}$ is the number of districts that share boundaries with the i-th one (i.e. its neighbouring districts). s_i is the unstructured spatial effect in each district modelled using an exchangeable prior, si ~ Normal (0, σ_{ε}^{2}).

γ_t is the structured temporal component assuming (i) an AR (1) structure (Diggle, et al., 2002a); (ii) stationarity of the data over time and (iii) all the observations are regularly spaced in time. τ_t is the unstructured temporal heterogeneity in each district modelled using $s_t \sim N(0, \sigma^{2}{}_{\gamma})$

Let y_{it} denote the number of outbreaks occurring in year t t(t=1,….T) in each district i (i=1,….I). It is assumed that the number of outbreaks y_{it} for district i, in year t, has a Poisson distribution with parameters μ_{it} and probability π_{it} with a log link, where linear predictor μ_{it} decomposes additively into time and space dependent effects.

Extra-Poisson variability as discussed earlier can be explained by inclusion of spatial and temporal structured dependency in space and time respectively (Eq 1). When there is presence of extra-Poisson variability in the data even after accounting for spatial and temporal autocorrelation, additional parameters are required to account for this residual variability. The residual spatio-temporal variability can be accounted for by including certain interaction terms between the structured and unstructured components in both space and time respectively. These additional terms can be modelled by including either of four possible types of interactions. Thus, addition of any or all of the four types of interaction term leads to (up to) five parameters in the model to account for the extra-Poisson variability. The first four terms are spatial structured & unstructured heterogeneity and temporal structured & unstructured heterogeneity. The fifth term will be either of the four types of interactions discussed below.

The interaction parameter δ_{it} (i=1,….n, t=1,….T) is added to equation (1) to specify either of the four types of interaction. δ_{it} is assumed to follow a Gaussian distribution with precision matrix $\lambda_{\delta} K_{\delta}$. K_{δ} is specified as the Kronecker product of the structure matrices of the structured and unstructured (of spatial and temporal) components which interact. The Kronecker product is denoted by $\otimes$, which is an operation on two matrices of arbitrary size and it is different from routine matrix multiplication.

Type-1 interaction (independent interaction): In this type of interaction (Eq 2), in addition to the spatial and temporal (structured and unstructured) heterogeneities, the additional term is the interaction term between the unstructured spatial and temporal components. In this study, it means that although there is presence of spatial dependency and temporal dependency, there are errors which are not explained by these two terms. The residual errors are explained by the interaction between some omitted risk factors which are specific to each year and district without any spatial or temporal structure respectively. This can happen when there are unobserved covariates in each district and year, and they do not have any spatial and temporal structure. The omitted factors may be certain farm level factors which vary over space and time. The independent factors can be anything which has not been included in the analysis.

$$\log \mu_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_n x_{nit} + \varepsilon_i + s_i + \gamma_t + \tau_t + \delta_{it} \tag{2}$$

$$\delta_{it} = \text{type - 1 space - time interaction}$$

Here the two unstructured main effects ε_i (temporal) and τ_t (spatial) interact.

Type-2 interaction: In this type of interaction (Eq 3), the temporal dependency term interacts with the unstructured spatial component term. This means that the outbreaks are not only dependent on the previous year's outbreaks but this temporal dependency also interacts with some unobserved district level covariates.

$$\log \mu_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_n x_{nit} + \varepsilon_i + s_i + \gamma_t + \tau_t + \delta_{it} \tag{3}$$

$$\delta_{it} = \text{type - 2 space - time interaction}$$

Here the structured temporal main effects γ_t and unstructured spatial component ε_i interact.

Type-3 interaction: The detection of type-3 interaction (Eq 4) implies that the spatial structure interacts with the unstructured temporal component, i.e. the outbreaks spread to nearby districts and interact with missing covariates in the time domain modelled by the temporal unstructured heterogeneity term.

$$\log \mu_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_n x_{nit} + \varepsilon_i + s_i + \gamma_t + \tau_t + \delta_{it} \tag{4}$$

$$\delta_{it} = \text{type - 3 space - time interaction}$$

Here the structured spatial component s_i interact with the unstructured temporal component τ_t

Type-4 interaction: In this type of interaction (Eq 5), the spatial and temporal structures interact. There is spread of outbreaks between districts and this interacts with the temporal dependence of the outbreaks. This type of interaction is relevant to daily/weekly data but may not be important with yearly aggregated data because disease in one year at one place may not spread next year to neighbouring places and this may happen in short time (days or weeks).

$$\log \mu_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_n x_{nit} + \varepsilon_i + s_i + \gamma_t + \tau_t + \delta_{it} \qquad (5)$$

$$\delta_{it} = \text{type - 4 space - time interaction}$$

Here the both the structured spatial (s_i) and temporal (γ_t) terms interact with each other.

The best model identified from the combination of individual categories of predictor model was then tested for the four different types of interactions as discussed earlier to capture any residual extra-Poisson variability. The effect of the interaction terms is null if there is no residual extra-Poisson variability after accounting for spatial and temporal main effects independently (Knorr-Held, 1999). The interaction terms are always added in at the last stage of model building, to capture the extra-Poisson variability and to account for missing covariates with either of the four types of interaction (Schrödle and Held 2011).

The best model was used to make predictions on the out of fit data (2008-2009). The predictive power of the model was assessed by calculating RMSE (Root Mean Square Error) between observed and predicted bluetongue outbreaks on the out of fit data.

Model with Negative binomial distribution

Finally a negative binomial distribution was fitted by including an over dispersion parameter to check whether the extra-Poisson variability is due to overdispersion in the data and whether it improves the predictive performance of the model (Eq 6) with spatial and temporal random effects.

$$\mathrm{Pr}_{\mathrm{k,\,p}}(\mathrm{y}) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}\, p^{k}(1-p)^{y}$$

$$\text{Probabilit y } 0 < p \leq 1$$

$$k \in R,$$

$$k > 0, \Gamma(n) = (n-1)!\,\text{denotes Gamma function}$$

$$\mu = \frac{k(1-p)}{p}$$

$$\text{Mean and variance are given by} \qquad (6)$$

$$\mu = \frac{k(1-p)}{p}$$

$$\sigma^{2} = \mu + \mu^{2} / k = \frac{k(1-p)}{p^{2}}$$

$$Y_{it} \mid \mu_{it} \sim \text{negative binomial}\,(\mu_{it,}\,\sigma^{2})$$

$$\log \mu_{it} = \beta_{0} + \beta_{1}x_{1it} + \beta_{2}x_{2it} + \beta_{n}x_{nit} + \varepsilon_{i} + s_{i} + \gamma_{t} + s_{t}$$

$$\sigma^{2} - \text{variance}$$

5.2.4 Model building

All possible combinations of models with structured and unstructured components of spatial and temporal effects (four components in all; two for spatial and two for temporal) were first fitted within land-cover classes (1024 combinations), climate (64 combinations) and host categories (128 combinations) to identify the best models within these categories. The best model was identified within each category as the model with the lowest Deviance Information Criteria (DIC) (Spiegelhalter et al., 2002). Once the best model in each category had been identified, all possible model combinations (65536) of the best selected predictors were fitted and the best combined category model again was identified using DIC (Fig 5.1).

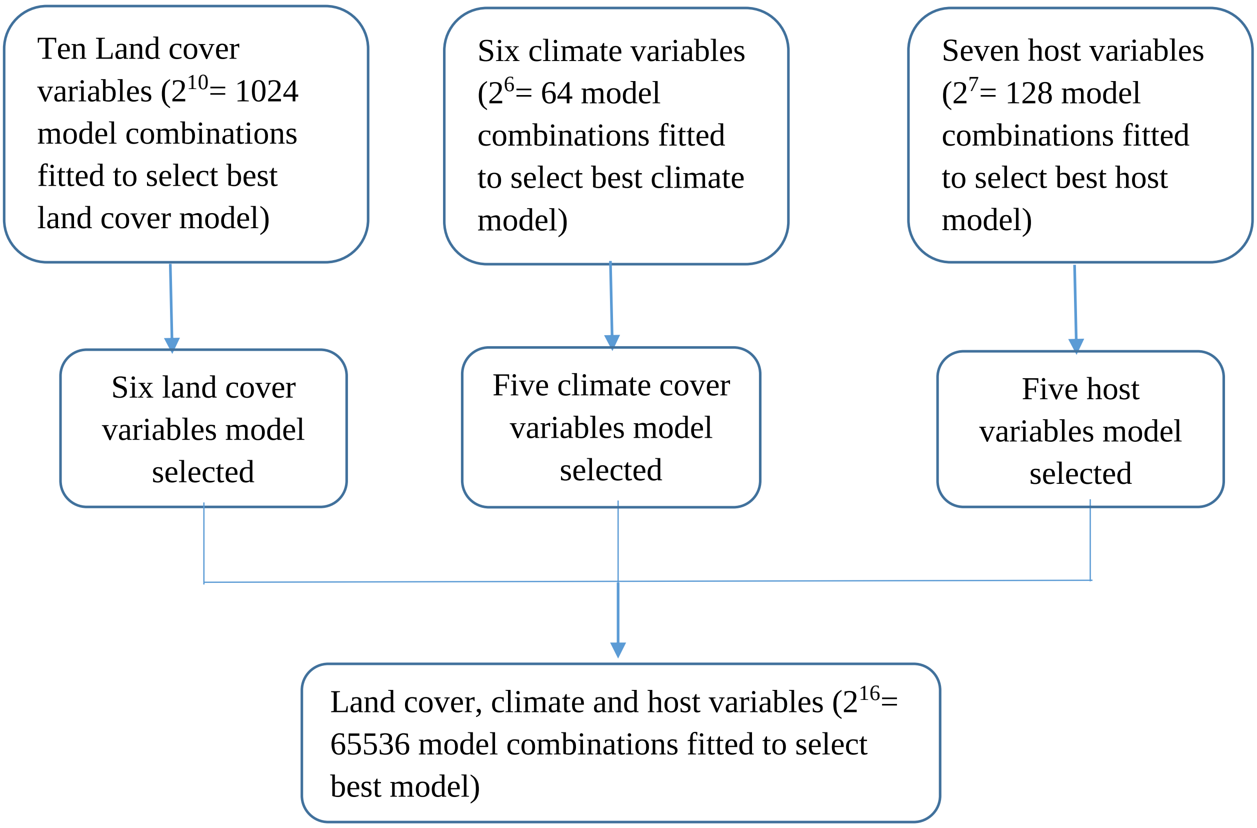

Ten Land cover variables (2^{10}= 1024 model combinations fitted to select best land cover model)
Six climate variables (2^6= 64 model combinations fitted to select best climate model)
Seven host variables (2^7= 128 model combinations fitted to select best host model)
Six land cover variables model selected
Five climate cover variables model selected
Five host variables model selected
Land cover, climate and host variables (2^{16}= 65536 model combinations fitted to select best model)

Figure 5.1: Variable selection procedure: Best variables were selected from individual category (land cover, climate and host) and the final model was selected by fitting different combinations of selected predictors.

5.3 Results

5.3.1 Spatio-temporal patterns in bluetongue outbreaks

The space-time multi panel plot shows that the BTV outbreaks occur every year (Fig. 5.2) and 60 out of a total of 80 districts reported BTV outbreaks on one or more occasions over the 18 year period of the records. The maximum annual number of outbreaks in a single year (2858) in all the districts was reported in 1998 followed by 1229 outbreaks in 2005. The time series plots for different districts of South India (Fig. 5.3) show that outbreaks occur with varying severity in different districts, with regular outbreaks occurring in some districts but not others.

5.3.2 Comparison of space-time models

Comparison of different models with and without adding spatial and temporal random effects within individual categories of models (climate, host and land cover) resulted in better performance of models with spatial and temporal random effects compared to covariate only models (Table 5.1). The model with covariates only was the worst performing model in all the individual categories of models. The model (climate) with covariates and spatial & temporal random effects had a lower DIC (DIC=9927) than models containing either covariates alone (DIC=29715) or by considering only spatial & temporal random effects and not including covariates (DIC= 10320) (Table 5.1).

Within the individual category models, the climate model outperformed the host and land cover models, with the latter the worst of the three (Table 5.1). The difference between the climate and host model is significant with a drop of 379 DIC units (McCarthy, 2007) and that between the climate and land cover models is slightly greater, with a drop of 386 DIC units. However, the combined model performed better than the individual category models, with a further drop of 8 DIC units.

The type-1 interaction model was the most parsimonious one among the four types of interaction considered. The best model with type-1 interaction performed better than the best model without interaction (Table 5.3).

The combined model with spatial and temporal random effects shows that the bluetongue outbreaks in each district and year are significantly and negatively associated with annual precipitation at lag 1 and annual mean maximum temperature at lag 0, and are positively associated with precipitation at lag 0 and maximum temperature at lag 1 (Table 5.2). Densities of non-descript sheep and exotic & crossbred breed of sheep are significant and positively associated with bluetongue outbreaks. None of the land cover classes was significant in the final model.

The predicted number of outbreaks using the best model shows excellent correspondence with the observed number of outbreaks (with a correlation co-efficient of r =0.8357 for the climate-only model, r= 0.8131 for the host-only model and r=0.8130 for the land cover-only model) with a BYM model (including covariates and spatio-temporal random effects) (Fig. 5.4). However, the correlation co-efficient with only covariates and no spatial and temporal random effects was lower (of r =0.2173 for the climate-only model, r= 0.2289 for the host-only model and r=0.1930 for the land cover-only model).

Comparison of the best model with and without interaction on the basis of its predictive performance on out of fit data using root mean square error resulted in a better model without interaction in both the years (Table 5.3). The models with and without interaction performed better on the out of fit data for the year 2008 than for the year 2009.

The fit of the model on training set (average of 1992-2007) shows good correspondence with the observed BTV outbreaks (Fig. 5.4). Comparison of the observed and predicted bluetongue outbreaks for the year 2008 and 2009 is shown in Fig. 5.5. The bluetongue outbreaks are predicted in districts of all the three states, but the observed outbreaks are shown only in a few districts of Karnataka.

The mean co-efficients of the fixed and random effects of the model with negative binomial distribution (Table 5.4) shows that only non-descript sheep and exotic & crossbred sheep are positively and significantly associated with BTV outbreaks. Comparison of the DIC between models with and without type-1 interaction and the correlation co-efficient between observed and predicted outbreaks using the negative binomial model is shown in Table 5.5.

Comparison between the DICs of the model with Poisson structure and the Negative binomial (NB) structure shows the NB model has a lower DIC than the model with Poisson errors, but has a poorer correlation co-efficient (r^2=0.43 versus r^2=0.83) between observed and predicted BTV outbreaks. The NB model performance is worsened (higher DIC value) when there is inclusion of type-1 interaction and also there is not much improvement in the correlation co-efficient from the model without type-1 interaction.

Wavelet analysis of the annual rainfall time series of three states (Fig. 5.6) shows a significant six to eight year periodicity but for very variable periods of time, longer in Tamil Nadu (c. 20 years, Fig. 5.6D) than elsewhere. This periodicity is not significant in any dataset before year 20 (year 35 for Andhra Pradesh and Karnataka) and does not coincide with any significant periodicity in the SST data (Fig. 5.6A) which itself shows brief significant periodicities of three to five years only and again only in the latter half of the time series. Thus, even if there is only temporary significance of periodicity in the state level rainfall data it seems unlikely that the cause of this periodicity is related to the SST in the eastern Pacific region. Correlation between the dominant frequencies of SST and annual rainfall in the three states (Fig. 5.7) shows a very brief (< five years) significant correlation at a four year periodicity from about year 15 (1964), but only in Andhra Pradesh. This state and Karnataka, but not Tamil Nadu, shows a strong correlation, again only temporary, at a periodicity of about six to eight years from the beginning of the dataset but fading from about year 20/25. Half or more of these periods are inside the cone of influence, COI, (i.e. should not be considered significant). Finally there is another short period of significance at a periodicity of four to seven years from about year 40 in Karnataka; again most of this falls inside the COI. There is no sign of a similar periodicity in the other two states. Tamil Nadu appears to be devoid of significant correlations, except briefly, at the start of the time series (Fig. 5.7C). This is also the case for the wavelet coherence analysis of South-West monsoon rainfall in the three states and El Nino (Fig. 5.8). There are virtually no areas of significance outside the cones of influence for any state.

The situation for the North East monsoon, however, is quite different (Fig. 5.9). In all three states there is a strong coherence between this variable and El Nino for the years 15 to 40 (i.e. years 1964 to 1989). Interestingly, the strongest correlation is for periodicities of 3 to 4 years initially, rising to 4 to 6 years by the end of this interval (year 40, 1989), continuing (non-significantly) to the end of the recording period (year 51, 2000). There are also signs of a strong correlation at a periodicity of 8 years for the first 15 years in each state, but this is mostly inside the cone of influence. There are few signs of this periodicity after year 15, when the 3 to 4 year periodicity 'takes over'.

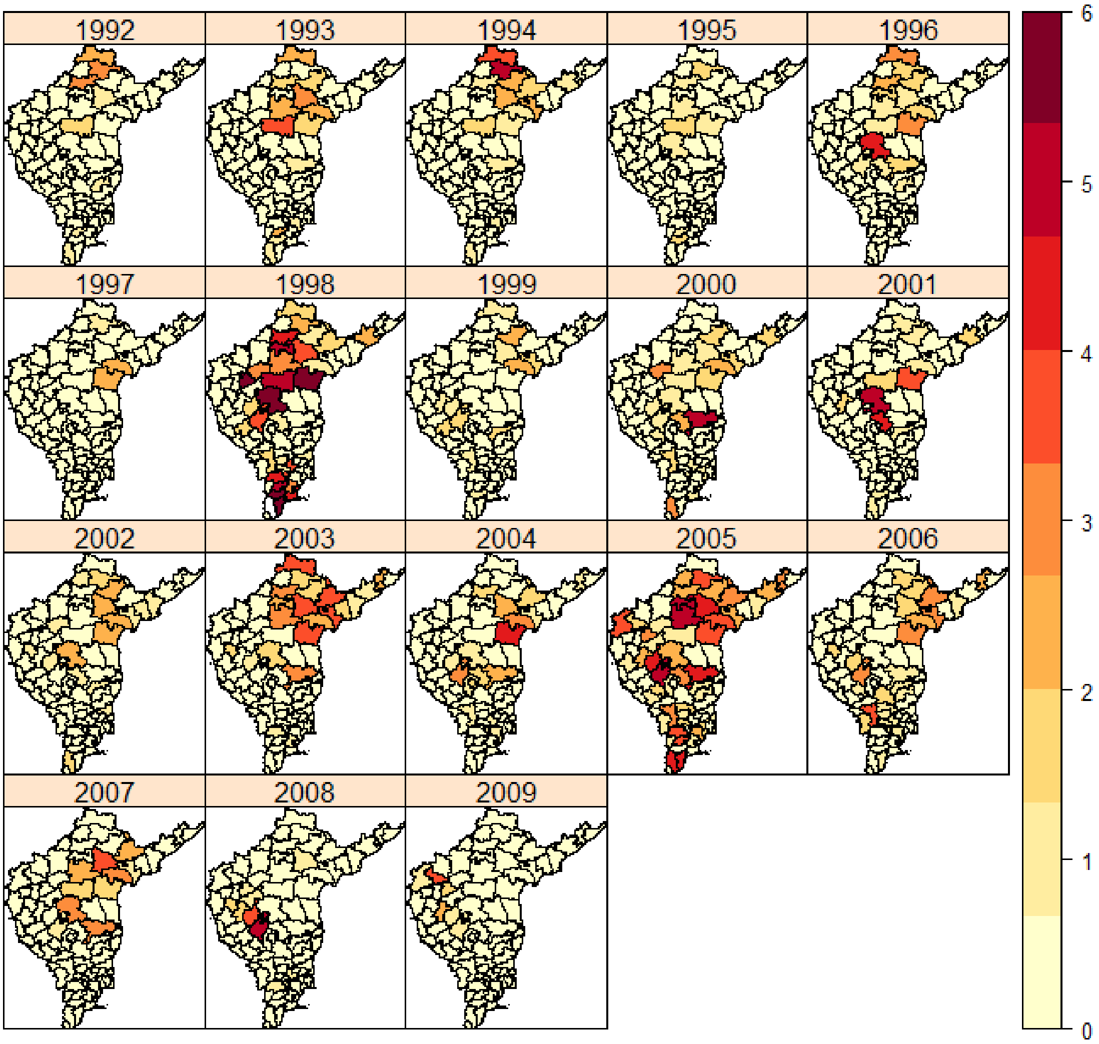


Figure 5.2: Bluetongue outbreaks in South India (1992-2009).The outbreak data were log.$_e$ transformed.

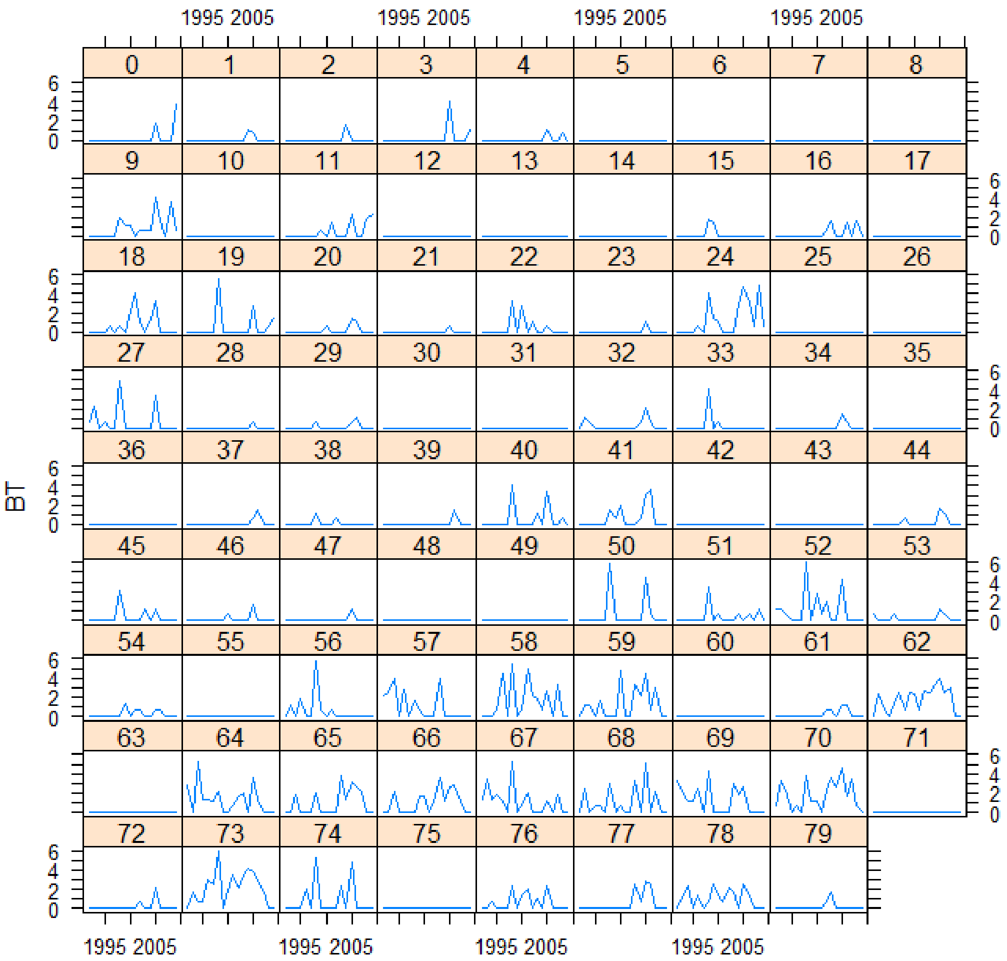


Figure 5.3: Bluetongue outbreaks from 1995 to 2006 in different districts of south India (1-79).The X-axis indicates different years and Y-axis indicates log$_e$ (natural logarithm) transformed BTV outbreaks.

Model	Predictors in the model	DIC (spatial and temporal random effects)	DIC (covariates only)	Correlation co-efficient (with spatial and temporal random effects)	Correlation co-efficient (covariates only)
Climate	Maximum temperature at 0, 1 &2 lag and precipitation at 0 & 1	9927.53	29715.75	0.8357	0.2173
Host	Non-descript sheep, exotic & crossbred sheep, Indigenous cattle, Crossbred cattle and Buffalo	10306.99	28081.27	0.8131	0.2289
land cover	Post-flooding or irrigated croplands, rain fed croplands, Mosaic croplands(50-70%)/Vegetation(grassland, shrub land, forest) (20-50%), closed to open (>15%) broadleaved evergreen and/or semi-deciduous forest(>5m), closed (>40%) broadleaved deciduous forest (>5m), closed (>40%) needle leaved evergreen forest (>5m)	10313.39	30378.19	0.8130	0.1930

Table 5.1: Deviance information criterion (DIC) and correlation co-efficient for the best models, where predictors are drawn from a single category of environmental predictors. Comparisons between models with spatial & temporal random effects and covariates, with covariate only models.

Variable	**Mean (sd)**	**Credible interval (CI)**
Intercept	-45.1435 (5.2210)	-55.7594(-35.2500)
Precipitation at lag 0	0.0009 (0.0001)	**0.0006, 0.0011**
Precipitation at lag 1	-0.0004 (0.0001)	**-0.0006 , -0.0001**
Maximum temperature at lag 0	-1.79 (0.14)	**-2.08 , -1.50**
Maximum temperature at lag 1	2.65 (0.16)	**2.34 , 2.98**
Non-descript sheep	2.82 (0.52)	**1.82, 3.90**
Exotic and crossbred sheep	0.61 (0.20)	**0.21, 1.01**
Mosaic croplands(50-70%)/Vegetation(grassland, shrub land, forest)	-13.19 (8.06)	-29.47, 2.30
closed (>40%) broadleaved deciduous forest (>5m)	2.84 (13.56)	-23.87 , 29.54
closed (>40%) needle leaved evergreen forest (>5m)	1.38 (31.55)	-60.56, 63.29
Random effects		
Spatial unstructured random effects	1.77(4.10)	**0.1095, 2.6944**
Spatial structured random effects	1931.56 (1877.46)	**143.23, 6859.3**
Temporal structured random effects	18234.55 (18140.90)	**1280.7313, 66064.10**
Temporal unstructured random effects	0.4647 (0.161)	**0.2143, 8.371**

Table 5.2: Mean coefficient values along with their standard deviation(sd) and credible intervals for fixed effects of environmental predictors with spatial and temporal effects (structured and unstructured components)which describe sum of annual number of bluetongue outbreaks per district (values in bold indicate that they are significant because the CIs do not bridge zero).

Model	**DIC**	**pD**	**r^2 on training data**	**r^2 on test data**		**Root mean square error**	
				2008	2009	2008	2009
Best model without type-1 interaction	9919.04	85.98	0.83	0.19	- 0.031	13.66	17.99
Best model with type-1 interaction	1940.31	420.72	0.99	0.35	-0.03	60.49	56.25

Table 5.3: best model with and without interaction along with their DIC's, pD (effective number of parameters), and the squared correlation- co-efficient (r2) between observed and predicted data, and Root mean square error on the test data

Variable	**Mean (sd)**	**Credible interval (CI)**
Intercept	-21.10 (6.91)	-34.95, -7.69
Precipitation at lag 0	-0.0011 (0.0006)	-0.0023, 0.0001
Precipitation at lag 1	0.0003 (0.0007)	-0.0010, 0.0016
Maximum temperature at lag 0	-0.85 (0.64)	-2.09, 0.42
Maximum temperature at lag 1	0.98 (0.64)	-0.30, 2.22
Non-descript sheep	3.01 (0.47)	**2.12, 3.98**
Exotic and crossbred sheep	0.56 (0.17)	**0.23, 0.91**
Mosaic croplands(50-70%)/Vegetation(grassland, shrub land, forest)	-5.77 (7.01)	-19.94, 7.71
closed (>40%) broadleaved deciduous forest (>5m)	8.51(11.71)	-14.44, 31.68
closed (>40%) needle leaved evergreen forest (>5m)	-0.22 (31.53)	-62.13, 61.62
Random effects		
Over dispersion parameter	0.23 (2.06)	**0.19, 0.27**
Spatial unstructured random effects	0.30(8.25)	**0.17, 0.49**
Spatial structured random effects	1699.45(1.71)	**114.04, 6276.72**
Temporal structured random effects	0.42(1.63)	**0.16, 0.79**

Table 5.4: Mean coefficient values along with their standard deviation(sd) and credible intervals for fixed effects of environmental predictors with spatial and temporal effects (structured and unstructured components)which describe the sum of annual number of bluetongue outbreaks per district (values in bold indicate that they are significant because the CIs do not bridge zero) using a negative binomial distribution model.

Model	**DIC**	**pD**	**r^2 on training data**
Negative binomial model without type-1 interaction	2656.26	71.22	0.426
Negative binomial model with type-1 interaction	2662.82	68.91	0.423

Table 5.5: Negative binomial model with and without interaction along with their DIC's, pD (effective number of parameters), and the squared correlation- co-efficient (r2) between observed and predicted data.

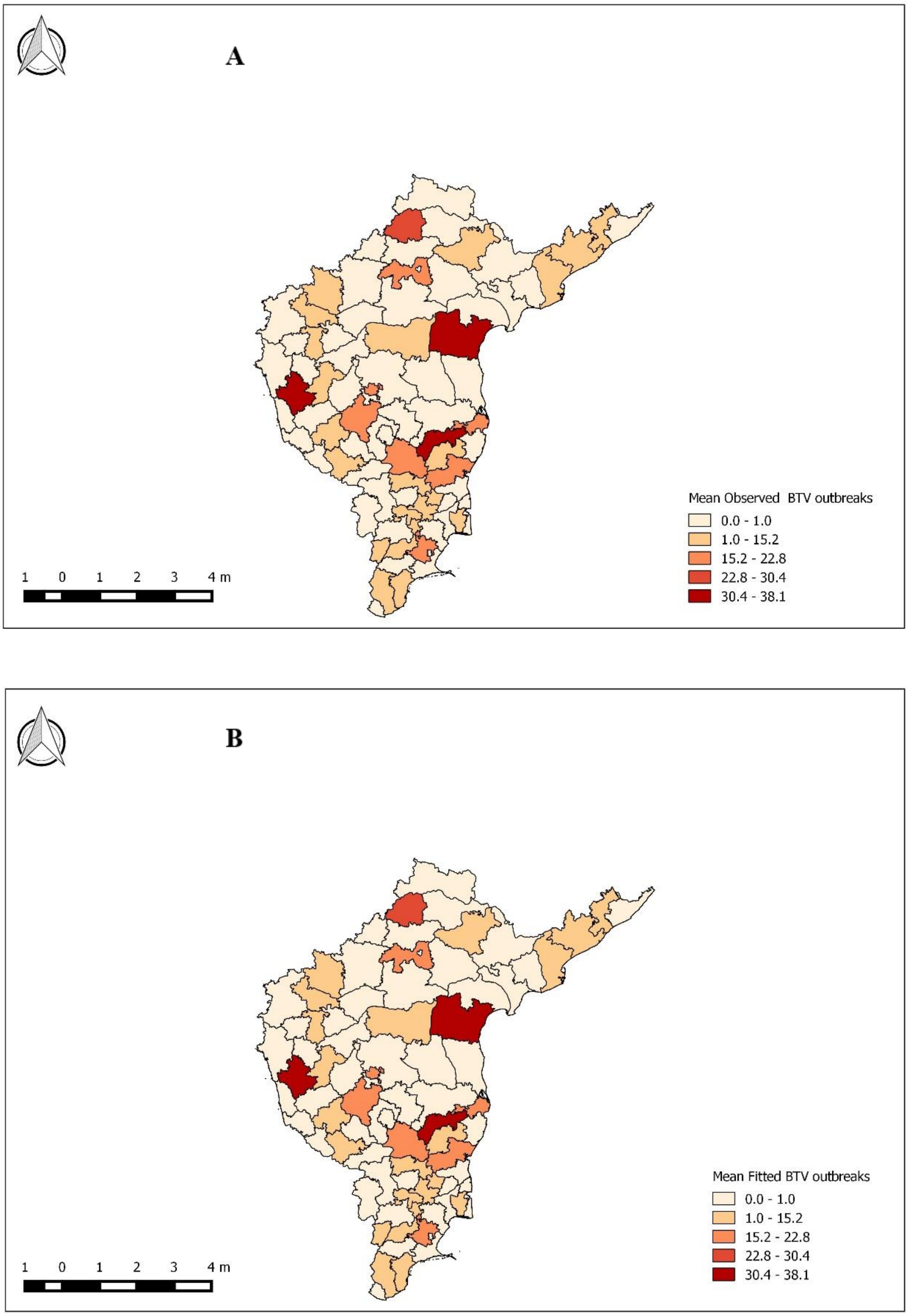
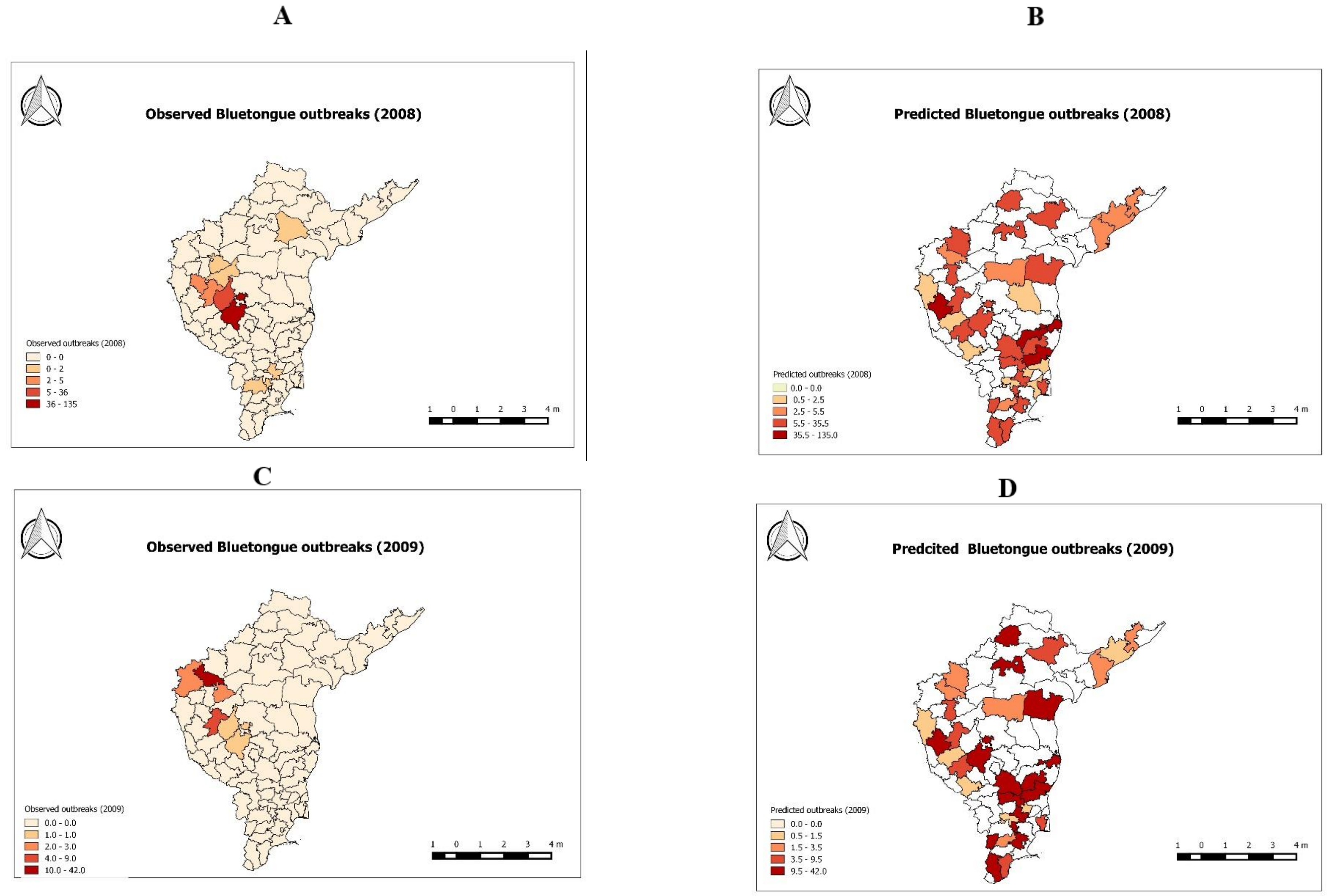


Figure 5.4: Comparison of the observed (A) average BTV outbreaks (1992-2007) with (B) fitted outbreaks using best model with spatial and temporal random effects.



A
Observed Bluetongue outbreaks (2008)
Observed outbreaks (2008)
0 - 0
0 - 2
2 - 5
5 - 36
36 - 135
1 0 1 2 3 4 m
B
Predicted Bluetongue outbreaks (2008)
Predicted outbreaks (2008)
0.0 - 0.0
0.5 - 2.5
2.5 - 5.5
5.5 - 35.5
35.5 - 135.0
1 0 1 2 3 4 m
C
Observed Bluetongue outbreaks (2009)
Observed outbreaks (2009)
0.0 - 0.0
1.0 - 1.0
2.0 - 3.0
4.0 - 9.0
10.0 - 42.0
1 0 1 2 3 4 m
D
Predcited Bluetongue outbreaks (2009)
Predicted outbreaks (2009)
0.0 - 0.0
0.5 - 1.5
1.5 - 3.5
3.5 - 9.5
9.5 - 42.0
1 0 1 2 3 4 m

Figure 5.5: observed and predicted bluetongue outbreaks using the model from best combination of variables with spatial and temporal random effects on the "out of fit" data for two years (2008 & 2009). (A) Observed bluetongue outbreaks for the year 2008 (B) Predicted outbreaks for the year 2009 (C) observed bluetongue outbreaks for the year 2009 and (D) observed bluetongue outbreaks for the year 2008.

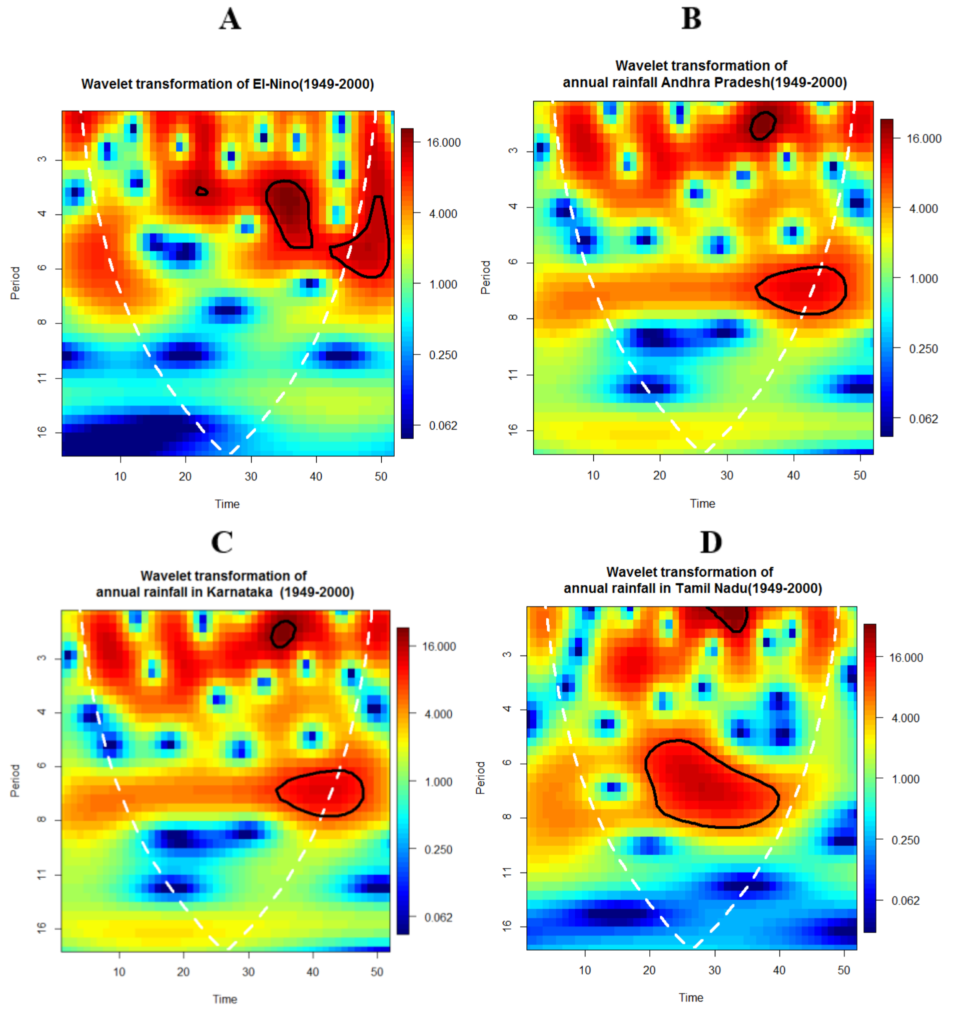
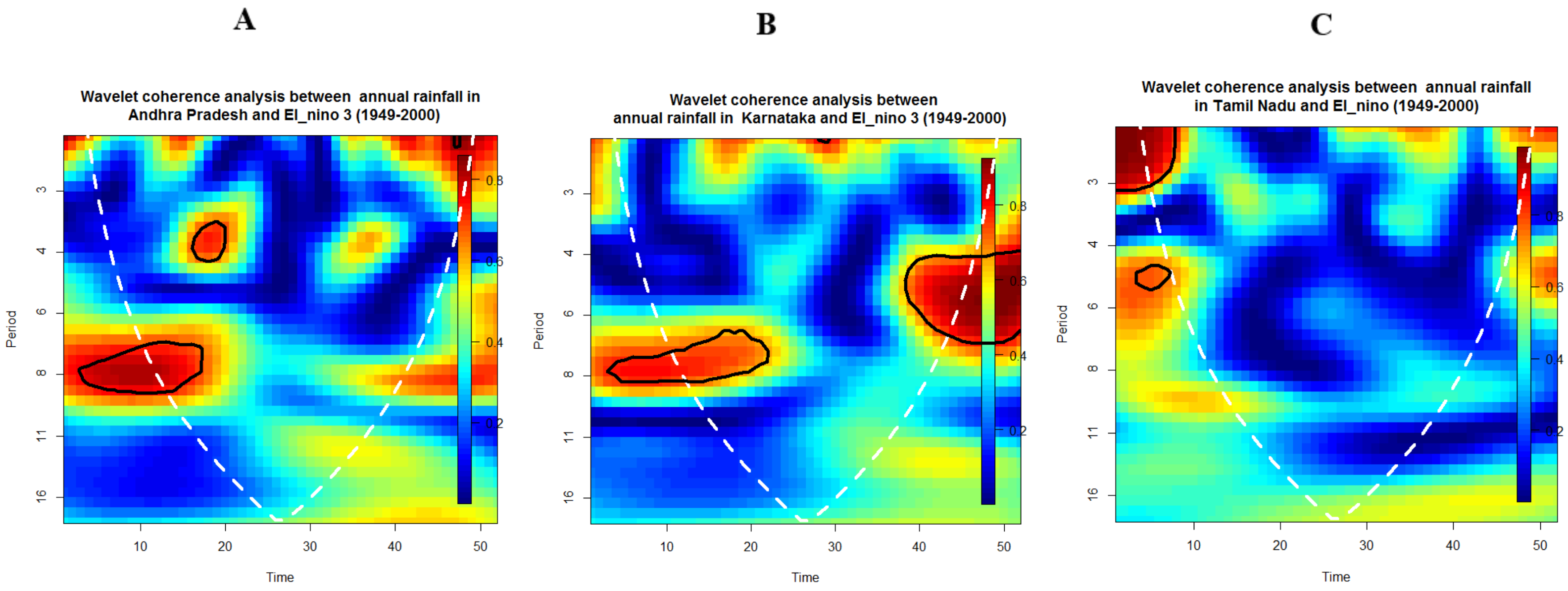


Figure 5.6: Wavelet power spectra of- the time series of annual rainfall in three Indian states & of the annual sea surface temperature in the eastern Pacific region. The white dotted line is the cone of influence indicating the region of time and frequency where the results are not influenced by the edges of the data and are therefore reliable (these areas are within the bullet shape formed by the two white lines and are said to be 'outside the cone of influence'; all other areas, physically outside the bullet shape, are 'within the cone of influence' of edge effects, and so are not reliable). The solid black line corresponds to the 95% confidence interval and the areas within this black solid line indicate significant variability at the corresponding periods (y-axis, in years) and time (x-axis, in years from 1949 (= Year 1). (A) Wavelet power spectrum of the El-Nino time series. (B) Wavelet power spectrum of the Andhra Pradesh annual monsoon rainfall time series. (C) Wavelet power spectrum of the Karnataka annual monsoon rainfall time series. (D) Wavelet power spectrum of the Tamil Nadu annual monsoon rainfall time series. The wavelet spectrum is shown with power increasing from blue to red colours. X-axis: time in years from the start of the time series (1949), Y-axis: periodicity, years.



A

Wavelet coherence analysis between annual rainfall in Andhra Pradesh and El_nino 3 (1949-2000)

B

Wavelet coherence analysis between annual rainfall in Karnataka and El_nino 3 (1949-2000)

C

Wavelet coherence analysis between annual rainfall in Tamil Nadu and El_nino (1949-2000)

Figure 5.7: Correlation (Cross-wavelet coherence) between dominant frequencies in the annual rainfall time series & annual sea surface temperature (El-Nino 3) time series.Wavelet power spectrum- The white dotted line is the cone of influence indicating the region of time and frequency where the results are not influenced by the edges of the data and are therefore reliable. The solid black line corresponds to the 95% confidence interval and the areas within this black solid line indicate significant variability at the corresponding periods and times. (A) Andhra Pradesh annual rainfall and El-nino 3. (B) Karnataka annual rainfall and El-nino 3. (C) Tamil Nadu annual rainfall and El-nino 3. Cross-wavelet coherence (correlation) and the wavelet spectrum is shown with power increasing from blue to red colours (scale is from zero to one with maximum correlation as one and no correlation as zero). X-axis: time in years from the start of the time series (1949), Y-axis: periodicity, years. The wavelet spectrum is shown with power increasing from blue to red colours.

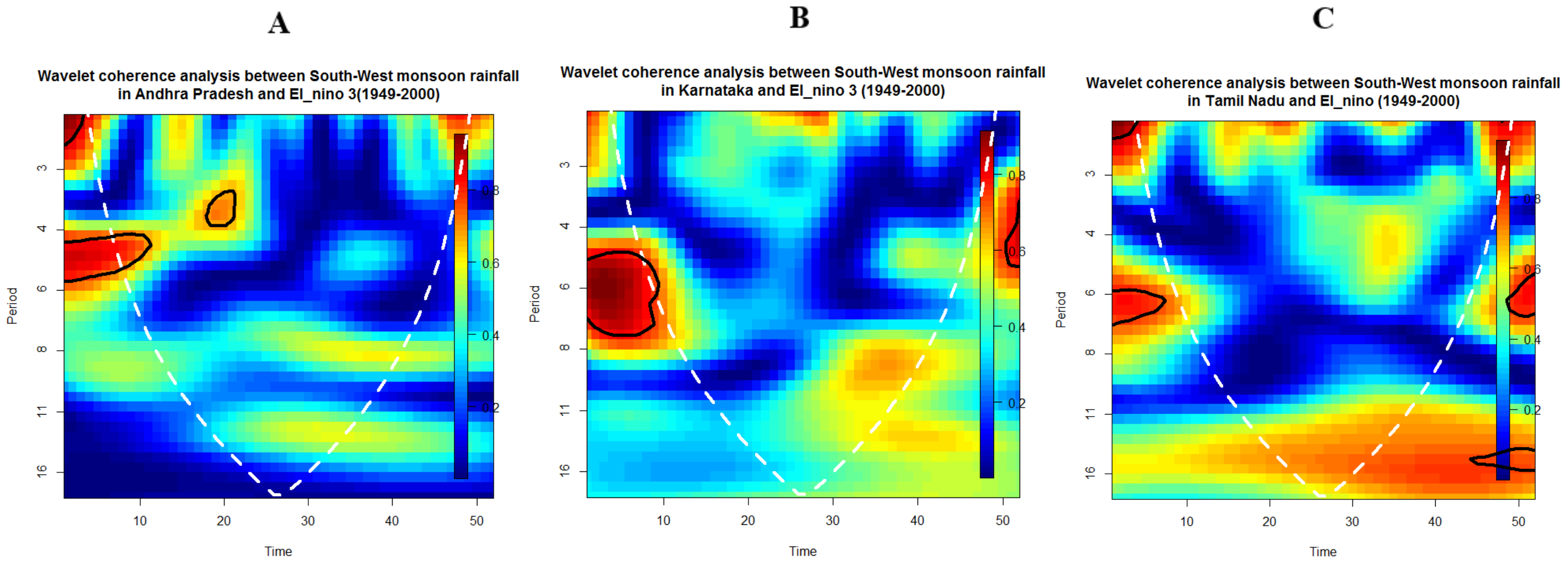


Figure 5.8: Correlation (Cross-wavelet coherence) between dominant frequencies in the South-West monsoon rainfall time series & annual sea surface temperature (El-Nino 3) time series. Wavelet power spectrum- The white dotted line is the cone of influence indicating the region of time and frequency where the results are not influenced by the edges of the data and are therefore reliable. The solid black line corresponds to the 95% confidence interval and the areas within this black solid line indicate significant variability at the corresponding periods and times. (A) Andhra Pradesh South-West monsoon rainfall and El-nino 3. (B) Karnataka South-West monsoon rainfall and El-nino 3. (C) Tamil Nadu South-West monsoon rainfall and El-nino 3. Cross-wavelet coherence (correlation) and the wavelet spectrum is shown with power increasing from blue to red colours (scale is from zero to one with maximum correlation as one and no correlation as zero). X-axis: time in years from the start of the time series (1949), Y-axis: periodicity, years. The wavelet spectrum is shown with power increasing from blue to red colours.

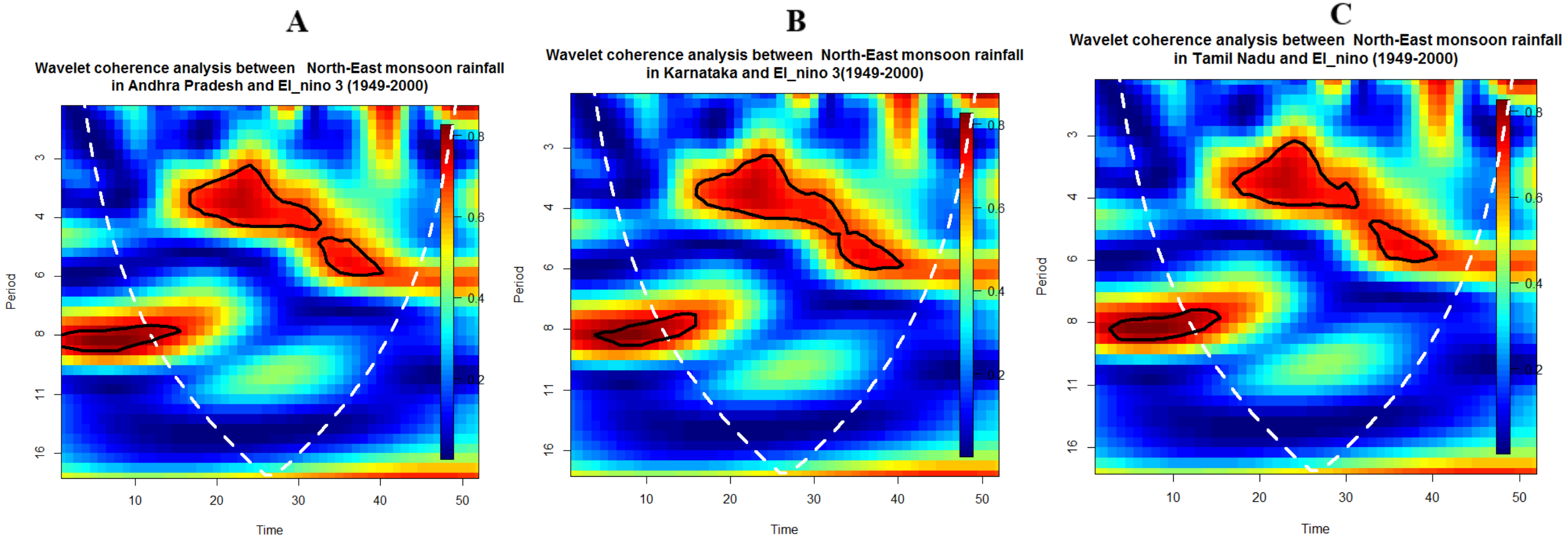


Figure 5.9: Correlation (Cross-wavelet coherence) between dominant intra-annual frequencies in the North-East monsoon rainfall time series & annual sea surface temperature (El-Nino 3) time series.Wavelet power spectrum- The white dotted line is the cone of influence indicating the region of time and frequency where the results are not influenced by the edges of the data and are therefore reliable. The solid black line corresponds to the 95% confidence interval and the areas within this black solid line indicate significant variability at the corresponding periods and times. (A) Andhra Pradesh North-East monsoon rainfall and El-nino 3. (B) Karnataka North-East monsoon rainfall and El-nino 3. (C) Tamil Nadu North-East monsoon rainfall and El-nino 3. Cross-wavelet coherence (correlation) and the wavelet spectrum is shown with power increasing from blue to red colours (scale is from zero to one with maximum correlation as one and no correlation as zero). X-axis: time in years, Y-axis: period. The wavelet spectrum is shown with power increasing from blue to red colours.

5.4 Discussion

The results of the space-time South India model indicate that the combination of a high maximum temperature and low rainfall in the year preceding an outbreak and a low maximum temperature and high rainfall in the year itself resulted in more outbreaks over the study period. The former conditions (high temperature and low rainfall in the preceding year) can alter the breeding sites of larvae. Drought seems to be important in the epidemiology of EHDV (Epizootic Haemorrhagic Disease Virus) (Dubay et al., 2004).

The positive and significant association of the non-descript sheep and exotic and crossbred breed of sheep with BTV outbreaks is consistent with the findings of the spatial analysis of bluetongue outbreaks in Andhra Pradesh and Karnataka (Chapter 4). Although similar land cover classes (post-flooding or irrigated croplands and rain fed croplands) as in the spatial analysis were also selected within the individual category of models in the present analyses, they were not selected in the combined best model, and none was significant in the final model. Therefore although land cover and host types were important in determining average BTV severity across years at the district level they do not seem to be important in determining the inter-annual variability of BTV outbreaks between districts. However, use of dynamic host and land cover variables (especially the rain fed and irrigated croplands) is required to rule out their role in describing inter-annual variability.

Instead models using just climate data outperformed host and land cover models in the individual category of models and also in the final model, in which host variables were also included. Thus, climate appears to play a dominant role in the inter-annual variability of BTV outbreaks in each district and substantiates the findings of the time series analyses at the state level (Chapter 3).

The variance explained by the individual category models without adding any spatial and temporal random effects is no greater than 20%. The variance explained by a climate model for BTV in Israel was also around 20% (Purse, Baylis, et al., 2004). The low variance explained in the present case may also be due to high, variable levels of immunity to the numerous (>20) circulating serotypes of BTV.

The best model with spatial and temporal structure (Fig. 5.4) predicts bluetongue outbreaks, albeit often at a very low level, in all districts of all three states in 2008 and 2009 (the 'out of fit' years), whereas the disease was actually recorded from only relatively few districts of Karnataka in those years. The correlation between observed and predicted outbreaks is better in 2008 (r^2= 0.35) than in 2009 (r^2=-0.03). Thus, whilst the introduced temporal structure (with an AR(1) component) could in theory have accounted for some of the effects of the possible high levels of immunity suggested in the previous paragraph, it appears that it was not doing so in the present instance. Either the effect is not there (i.e. immunity is not important) or it is much more complicated than is imagined in a relatively simple analysis (perhaps due to the interaction between multiple co-circulating serotypes of BTV, or due to vector numbers varying between years independently of host immune status, as suggested in the time series analysis chapter).

Wavelet coherence analysis (correlation) between El-Nino 3 and different annual and seasonal rainfall variables from different states resulted in strong correlation for about half the total recording period between North-East monsoon rainfall and sea surface temperature for all three states for periodicities of about three to six years (Fig. 5.9). There was significant correlation of the El-Nino-3 index with North-East monsoon rainfall in Tamil Nadu (r^2=0.44, $p < 0.005$), South interior Karnataka region (r^2=0.50, $p < 0.005$) and a non-significant correlation for coastal Andhra Pradesh (Zubair & Ropelewski, 2006). Significant periodicities of between two and seven years were also found in the El-Nino data but these, too, are transient, and occur only in the second half of the recorded period (Fig. 5.6A and Hales et al., 1999; Kovats, 2000). The timing of the North-East monsoon rainfall (Oct-Dec) coincides with the major proportion of the BTV outbreaks in Tamil Nadu but outbreaks start in the month of August and decline from October onwards in Andhra Pradesh as shown in the monthly time series analysis (Chapter 3). However, Karnataka is influenced by both South-West and North-East monsoon and therefore the outbreaks are observed in both the seasons but in different geographical regions of the state. Establishing links between El Nino and local rainfall (Fig. 5.9), however, brings us no nearer to solving the problem of BTV outbreaks and can only eventually contribute to a DEWS if and when the link between BTV and rainfall is formally established.

Inter-annual variability in climate influenced by El-Nino may affect the socio-economic conditions of the farmers dependent on monsoon for fodder. Drought conditions are also believed to increase midge-wildlife (Zebra; reservoir host) contact rates and thus further increase BTV transmission (Baylis et al., 1999b). The chance of contact with wild-life reservoir hosts (deer and other wild ruminants) is higher the more mobile are the sheep populations. Such sheep movement is more common in drought years, due to the scarcity of fodder. Drought also promotes long distance migration of sheep farmers to distant places, which might lead to a higher risk of contact with wild ruminants (some species of deer) which are sub-clinically infected reservoirs of bluetongue in Europe (Garcia-Saenz et al., 2014; Ruiz-Fons et al., 2008). Drought conditions will also increase contact rates between midges and livestock due to congregation of animals near limited water resources.

Incorporating spatial and temporal structure and the type -1 interaction resulted in a better model than the model without any random effects (Table 5.1). There are many examples of Bayesian disease mapping accounting for spatial and temporal heterogeneity. In most of the examples a CAR prior was used to account for spatial structure, but the temporal prior was more like a random walk (Knorr-Held & Besag, 1998) or using regression B-splines (MacNab et al., 2004). Neglecting the spatial structure or temporal structure can lead to selection of variables which may not be important (Hoeting et al., 2006), and also to loss of predictive power (Wikle, 2002). Accounting for spatial and temporal structure also accounts for omitted predictor variables having spatial or temporal structure.

Likewise, accounting for interaction (type-1 in our study) is important to prevent bias in estimation of parameters (Blangiardo et al., 2013).

The best combined model (including both climate and host significant variables as well as spatial and temporal structured and unstructured random events) increased these r-squared values to up to 99% on the training data (model with type-1 interaction). Unfortunately, however, these high predictive accuracies were not shown with the ‘out-of-fit’ data for 2008 and 2009, where explained variances fell to 35% and only 3% respectively. Model fit to the training set data appears in this case to be a very poor indicator of out of fit predictive performance. The reasons for this urgently need to be investigated but, until we have answers to some of the questions raised here, the likelihood of an accurate DEWS for BTV in India seems very low.

Chapter 6

Village level spatial analysis of bluetongue cases using Bayesian Network modelling and Bayesian model based geostatistics

6.1 Introduction

Bluetongue outbreaks have been recorded since 2004 in more than 25,000 (26,613) villages in Andhra Pradesh. In addition to the occurrence of outbreaks in villages, the numbers of animals affected and the numbers dying of bluetongue are also recorded. Models and analyses are required to help us to understand the role of different risk factors in determining the severity of BTV cases at village level. There is a great deal of geographical variability in bluetongue cases in Andhra Pradesh and not all villages are equally affected by the disease. Host and land cover predictors were shown to be important in discriminating between less and more severely affected districts in the spatial analysis of Chapter 4. Establishing the relationships between bluetongue cases and climate, host, and land cover variables at the village level should refine this understanding at a finer spatial scale.

Quantifying the role of different predictors in governing spatial variability in disease severity between villages is important in control and management of bluetongue. There are more than eighty thousand villages in South India and more than 50,000 of them have sheep. Currently there is no vaccination programme to control bluetongue, but pentavalent vaccine technology developed under the All India Network Programme (AINP) has recently been transferred to commercial vaccine manufacturers that will enable large scale production of a vaccine. Currently vaccination for other infectious diseases in South India is carried out by field veterinarians in rural villages (Heffernan et al., 2011) and all the villages are covered by vaccinating against major diseases. Vaccination is not targeted in any way (e.g. to villages historically at higher risk than others). Given the huge number of villages and limited veterinary personnel to cater for the needs of livestock farmers, fine scale village level risk maps would be very useful for estimating the approximate number of vaccination doses required, and for targeting vaccination and resources to high risk areas.

The present chapter focuses on identifying important variables in determining the severity of bluetongue cases and to develop a predictive model by accounting for spatial autocorrelation with two broad objectives:

1) To identify variables directly associated with bluetongue cases using Bayesian Network Model (BNM) analysis.
2) To use the variables identified in the BNM analysis to make predictions in unsampled villages using a Bayesian geostatistical model.

6.1.1Variable selection methods

In analyses of epidemiological data and risk factor identification a multitude of potential predictor variables must be reduced to a subset that is most strongly associated with the outbreak data. Analyses are usually correlative rather than causal but it is hoped that within the subset of predictor variables causal factors (or their proxies) are represented. It is the job of future research to confirm causal relationships suggested by correlation analyses. In reducing a large set of potential predictors to the smaller, useful subset, numerous variable selection procedures are available to the researcher, including significant p-values (Greenland, 1989), step-wise forward or backward procedures (Babyak, 2004), information criterion (AIC, AIC$_c$ and BIC) (Burnham & Anderson, 2004), LASSO methods (Tibshirani, 1997), least angle or penalized regression (Hesterberg et al., 2008), and all subsets approaches (Furnival & Wilson, 1974). Different variable selection methods with their advantages and disadvantages are discussed by (Miller, 2002). Recently, a review of the use of different variable selection methods in epidemiology found that 28% of the papers used prior knowledge to select variables, 20% used step wise selection procedures and a further 15% used changes in an accuracy metric (such as AIC or BIC) (Walter & Tiemeier, 2009). However, 35% of the publications did not describe the method in detail; variables were selected without any explanation for their choice (Walter & Tiemeier, 2009).

A common problem encountered in epidemiological analyses is that of correlation between predictor variables (in spatial or temporal domains). To remove obvious correlations such as these the predictor data may be first pre-processed using ordination techniques such as Principal Components Analysis (PCA) (Shapiro et al., 2005; Wade et al., 2008). Problems remain, however, of correlations between raw or pre-processed data of related types (for example, day and night-time Land Surface Temperature). These problems may be resolved by submitting the entire predictor data set to PCA, but the biological interpretation of each PC axis then becomes extremely difficult.

6.1.2 Bayesian Network Modelling

Considering the problem of multicollinearity and disadvantages of the other variable selection methods discussed above, BNM offers potential in identifying the important variables and their relationships. There is often some relationship or dependency between the sorts of predictor variables used in epidemiological studies and between some or all of these and the response variable, the number of cases. Whilst relationships may be nothing more than correlation (a and b are correlated; both might depend upon a hidden factor, c), dependencies imply causation (a causes b, but b does not cause a). Statistical analysis of epidemiological data rarely separates correlates from causes and it is left to the researcher, *post-hoc*, to make this distinction based on his/her experience, or further observations, interventions or experiments. Recently, Bayesian Network Modelling (BNM) has been employed to analyse multivariate data, to select only those variables that are in some way linked to each other and, in some cases, to identify potential causal pathways within the network (Lewis & Ward, 2013). BNM can be carried out in a number of ways; certain parts of the network can be defined in advance (when dependencies are known *a priori*) or BNM algorithms can suggest the most likely links between all the variables (both predictors and predicted) in the dataset (the more common scenario for epidemiological datasets). BNM proceeds by making each input variable a node and then seeks the links between the nodes on the basis of multi-variate Bayesian regression and GLM (Lewis and Ward 2013). The analysis determines the likelihood of each possible link in the network and the output is a graphical representation of the most likely links (arcs) given that particular dataset. Variables in the input dataset that are not linked to any other variables in the final network are omitted from the network diagram, and the strength and direction of the links in the network indicate the association between the nodes. Each node may have 1 to n parents (i.e. contributing directly to the node) and may connect to any number of offspring (child) nodes. It is tempting to think of parent nodes as implying causation between parent and offspring but this interpretation is only correct when the dataset is complete (i.e. when it contains all possible contributory variables). When it is not (which is usually the case with epidemiological datasets) causation cannot be certain because a hidden (i.e. unmeasured or unrecorded) variable may be determining both the parent and child nodes which may otherwise be conditionally independent of each other. Heckerman gives a salutary example of this effect (Heckerman, 1998). The number of parents for each variable to be tried is n-1 (n being the number of variables) but it can be restricted to fewer than that depending on the aims of the analyses and the computational resources available. In one example in the literature, the marginal likelihood did not improve beyond a five parent limit in a set of thirteen variables (Wilson et al., 2013).

Finally it should be stressed that there is no single ‘dependent’ variable in a BNM analysis (e.g. case numbers of a disease). All variables are treated equally, and the resulting network shows the links between them all. It is possible to exclude certain links within BNM (for example some nodes can be specified as having no parents, or others no child nodes) based on *a priori* knowledge about certain relationships.

Bayesian Network analysis has to date been applied in many fields such as neuroscience, bioinformatics, ecology and recently in epidemiological data analysis, to identify risk factors for child diarrhoea in Pakistan (Lewis & McCormick, 2012) and to explore associations between climate, farm management and the presence of tick species (*Ornithodorus. Erraticus)* on pig farms (Wilson et al., 2013). In this study (tick presence on pig farms), it was found that farm management practices were directly associated with the presence of ticks (four successive surveys over twelve year period) but that climate (temperature and precipitation) was not. In another study the association of weather with the occurrence of ten major diseases of pigs was studied (McCormick et al., 2013). Three pathologies were directly associated with temperature variables and all ten pathologies were related to at least two other pathologies each.

Intrinsic Spatial autocorrelation (SAC) can also be a problem with fine scale georeferenced point data, but conventional regression techniques such as linear models or generalized linear models employed in analysing infectious disease data fail to account for SAC. Fitting complex geostatistical structures within the BNM framework is computationally complex and time consuming, due to fitting of all possible networks (2^n combinations with single parent limit, where n is the number of variables).

Due to the robustness of BNM to identify the direct and indirect association between environmental variables and the advantage of fitting all possible combinations of networks, the present study applied BNM as an exploratory tool to identify the relationships between bluetongue cases, climate, land cover and host conditions in the village level Andhra Pradesh dataset (objective 1). Selected sets of important variables identified by the BNM were then used later in a Bayesian geostatistical regression model for making predictions of bluetongue cases in other areas (objective 2).

6.1.3 Bayesian geostatistical model

Geostatistical techniques are commonly applied to point- or polygon-referenced data in veterinary epidemiological studies (Biggeri et al., 2006). The different parameters of the semi-variogram model usually employed are shown in Fig. 6.1. The semi-variogram graph shows the spatial autocorrelation of the measured sample points. The distance at which the graph levels out is known as the range and thus the sample locations within the range are spatially autocorrelated. The y-axis value at which the semi-variogram graph attains the range is known as the sill. The nugget is the semi-variance at distance zero and is usually attributed to micro-scale variation, or measurement error. The choice of different semivariogram models (linear, exponential, spherical, Matern) is usually based on the fit to the semi variance (classical geostatistics) or the model likelihood (Diggle et al., 1998) to the observations. The latter is referred to as model-based geostatistics (Diggle et al., 2003) which can handle non-normal data. Bayesian geostatistical modelling is advantageous for complex hierarchical specification in the data and assigning prior distributions for parameters.

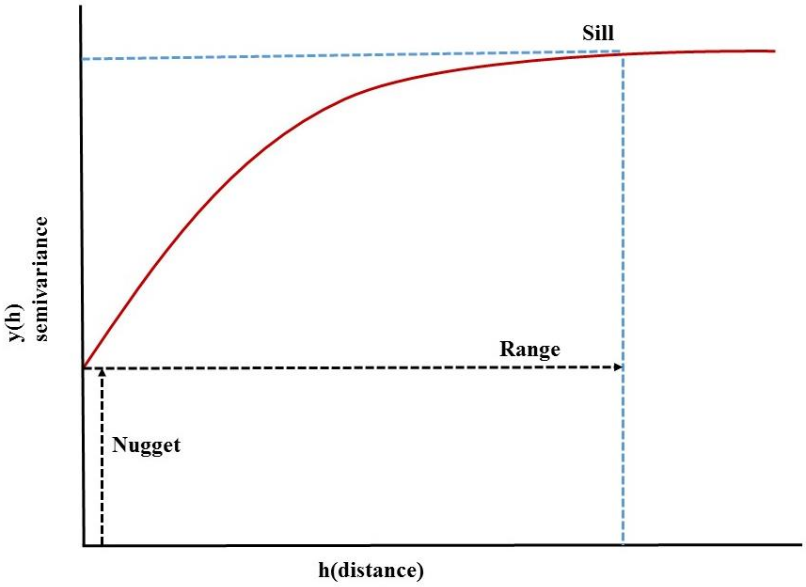


Figure 6.1: Semi-variogram showing different parameters; Nugget, range and sill

Problems with big datasets

Computing the spatial covariance matrix in geostatistics can be very difficult and time consuming as the number of observations (N) increases (Lasinio et al., 2013). Many methods have been proposed to estimate spatial covariance when N is large (Sun et al., 2012). Estimation using maximum likelihood approaches can be modified for a large number of observations by partitioning the data into clusters (Vecchia, 1988). These clusters are assumed to be conditionally independent and the likelihood is approximated. This method (Vecchia, 1988) was adapted using restricted maximum likelihood (Stein et al., 2004). In another method, the spatial process was represented using spectral processes (Fuentes, 2007) and approximating the likelihood. None of these methods is suited to non-stationary covariance situations (Banerjee & Fuentes, 2012).

Recent methods to solve problems with big datasets include approximation of the Gaussian field (GF) with a Gaussian Markov Random Field (GMRF)

(Rue & Held, 2005). This approximation of GF by GMRF makes the matrix sparse. There are computationally feasible algorithms for sparse matrices which can be used in the estimation of parameters, and recently (Rue et al., 2009) proposed the INLA (Integrated Nested Laplace Approximation) algorithm as an alternative to MCMC. Until recently it was not possible to fit geostatistical correlation structures with the INLA approach, but this was overcome by using an SPDE (Stochastic Partial Differential Equation) approximation of the GF by the GMRF (Lindgren et al., 2011). The stochastic partial differential equation creates links between Gaussian fields and Gaussian Markov random fields for faster estimation of spatial covariance. The SPDE approximation of the GF with GMRF is promising but requires pre-processing of the data to create a triangulation matrix (Lasinio et al., 2013).

In the methods discussed so far inference was either drawn based on maximum likelihood and their modifications or using INLA for Bayesian inference. Although Bayesian geostatistical models are advantageous over maximum likelihood based methods, computation of the spatial correlation structure using MCMC becomes complicated as the number of observations increases. MCMC computes inverses of the N*p by N*p (N is the number of observations and p the number of variables) covariance matrix for every iteration and is therefore computationally time consuming when hundreds and thousands of iterations are required. It may not be possible to run such calculations on personal computers. The magnitude of the problem increases as N increases.

Problems of computing the spatial covariance matrix of big datasets were overcome by using a class of models known as low-rank or reduced rank models (Banerjee et al., 2008) using a sub-sampling approach. The main assumption in the low-ranked approach or sub-sampling is that the spatial correlation structure at the observed sites can be summarized on a sub-sample, which is representative of the whole set of observations. The representative sets of locations are referred to as “knots”. (Kriegel et al., 2011). There are different methods for creating knots for sub-sampling. One-way of selecting knots is by use of different clustering algorithms like connectivity-based clustering (also known as hierarchical clustering) or centroid-based clustering. In connectivity-based clustering, points (locations of disease cases) are more related geographically to nearby points than the points which are far away. Different distance metrics like Euclidean distance, Manhattan distance, or maximum distance can be used depending on the type of data and study undertaken. The knots based approach is known as predictive process modelling (Banerjee & Fuentes, 2012) and computing covariance functions at the knots also addresses the issue of non-stationarity (directional trend). Non-stationarity in this study can arise due to distinct farm level practices in different villages, which may influence bluetongue transmission, and also due to movement of animals in different directions.

In this study the **spBayes** package (Finley et al., 2007) was used as it is designed for complex models and handles large spatial datasets within a Bayesian framework and can accommodate non-Gaussian dependent variables. The covariates (directly related ones only) selected in the

Bayesian Network Modelling were used in this analysis to quantify their role in governing the severity of bluetongue cases in the villages of Andhra Pradesh and test the model accuracy on ‘out-of-fit’ data.

6.2 Materials and methods

6.2.1 Bluetongue case data for Andhra Pradesh

The case data for bluetongue between 2004 and 2011 at village level were obtained from the department of Animal Husbandry, Andhra Pradesh, India. The department collates the number of monthly cases and deaths due to bluetongue from their wide network of field veterinarians and regional disease diagnostic laboratories. The bluetongue case data were summarised as the maximum number of cases per month occurring in a particular village from 2004-2011. The maximum number of cases was selected instead of other measures of summarising the data, in order to understand the role of different predictor variables in driving severity (maximum cases) in each village.

6.2.2 Climatic predictors

Monthly Rainfall Estimates (RFE) were obtained from the NOAA/Climate Prediction centre RFE 2.0. (Xie et al., 2002) as seven-year averaged values (2004-2010) available at 8 x 8 km spatial resolution. These were processed to calculate the average annual totals of South-West rainfall (June-Sept), North-East rainfall (Oct-Dec) and annual rainfall (Jan-Dec). The annual mean temperature layer between the years 1950-2000 was downloaded from Worldclim (Hijmans et al., 2005), available at 0.86x0.86km spatial resolution. The village level shape files for Andhra Pradesh were obtained under an individual license issued by the Survey of India and these were used to extract village centroids. The village level shape files (polygons) were used to extract rainfall variables and annual mean temperatures at the village level using the zonal statistics function in Arc Map 10.1 (ESRI, Inc., Redlands, CA, U.S.A.).

6.2.3 Land cover and topographical predictors

Digital elevation data were obtained from NASA's Shuttle Radar Topography Mission (SRTM) at ~90m resolution (Reuter et al., 2007). Aspect and slope were calculated at 90m resolution and averaged across the polygon for each village in Arc Map 10.1.

The areas of each village covered by different proportions of eight land-cover classes were extracted from the GlobCover land-cover map available at 300m resolution (Defourny et al., 2006) using the zonal statistics function in Arc Map 10.1 (ESRI, Inc., Redlands, CA,U.S.A.).

6.2.4 Livestock census data

Livestock censuses are carried out every five years and the 18th livestock census data (2007) were used in this study (http://www.dahd.nic.in/, accessed on 5th May 2014). Household surveys are conducted at village level involving a huge number of personnel for enumeration, supervision and compilation of the data at the all-India level. Densities of host species and individual sheep breeds, namely numbers of non-descript sheep, Deccani sheep, crossbred & exotic sheep, indigenous cattle and buffaloes, were extracted from the database of National Livestock census data and log-transformed to normalise the data. All the variables (except bluetongue cases with a Poisson distribution) were assumed to follow a normal distribution in a BNM analysis (Table 6.1). The variables can be allowed to follow other distributions and there is no pre-requisite in BNM analysis to follow certain distributions only. These host variables were selected based on a preliminary district level spatial analysis.

6.2.5 Bayesian network modelling

Bayesian network modelling is an extension of Bayesian graphical models routinely used in Bayesian analysis. In the single Bayesian graphical model, the structure of the graph and their dependencies are decided *a priori* and then the parameters are estimated. In BNM, however, the main objective is to derive a structure or graph which describes the relationships and interdependencies between variables. The difference lies in defining dependent and independent variables, which is pre-determined in the Bayesian graphical model, but is not specified in the BNM. The BNM algorithm searches the best graph among the different possible structures, based on the network score (log marginal likelihood) (Congdon, 2007).

The difference between a single Bayesian graphical model and BNM can be explained by using village level case data as an example. In a single Bayesian graphical model, the number of bluetongue cases can be considered as the dependent variable and the 20 environmental variables as independent (predictor) variables. The relationship between the dependent variable (e.g. cases) and the independent variables (e.g. environmental predictors) can be quantified using a Bayesian generalised linear model with Poisson errors. The probability function for Y is given by:

$$
\begin{aligned}
&\Pr(\mathrm{Y} = \mathrm{y}/\mu) = Poisson(\mu_i)\\
&\log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_n x_{ni} + v_i\\
&v_i \sim norm(o, tau.v)\\
&tau.v \sim gamma(mean, \mathrm{var})\\
&\beta_0 - \beta_5 \sim \mathrm{norm}\ (0, 0.001)
\end{aligned}
\qquad (1)
$$

Where β_0 is the intercept and β_1 to β_n are the coefficients for the fixed effects of the predictors x_{1i} to x_{ni}) in each place (village) *i*. and v_i is the unstructured spatial effect in each village, modelled using an exchangeable prior $v_i \sim$ Normal (0, var).

Equation 1 can be depicted by a graphical model

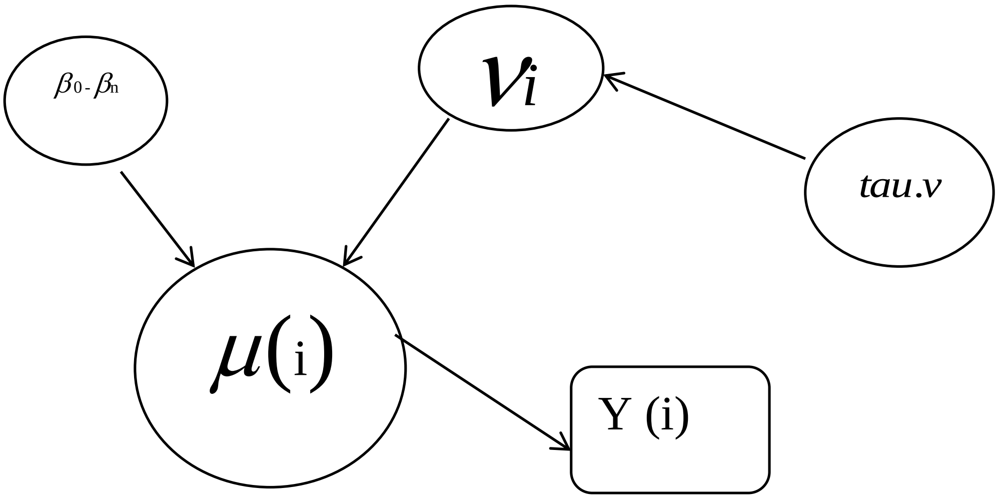


Figure 6.2: Depiction of a Bayesian Graphical model (DAG).The arrow indicates dependency and different parameters of the model are shown. The probability distribution of the dependent variable (Y (i)) is shown in a rectangle (dependent variable) and all the other parameters are shown in ellipses.

The above graph is referred to as a Directed Acyclic Graph (DAG) in Bayesian modelling. The cycle in the graph should not loop back, hence its ‘acyclic’ description (for example Y is dependent on μ but not *vice-versa*). This (DAG) terminology is commonly used in computer science and mathematics.

BNM is quite different from the Bayesian graphical model in that neither its structure nor its dependencies are decided *a priori*. Instead, the structure (DAG) is identified by fitting different relationships between different combinations of variables and examining which structure best describes the available data. None of the variables is considered as dependent or independent, but the best DAG identifies the links between the variables. As emphasised in the introduction to this Chapter, however, dependencies (implying causation) can only be inferred from a BNM when the input variable data set includes all possible contributory variables and omits no important (hidden) ones.

The BNM was fitted to the Andhra Pradesh dataset using the **abn** package in R. An uninformative structural prior was used, meaning that all the structures have equal chances of being selected in the final DAG, and uninformative Gaussian priors with mean zero and variance 1000 were assumed for the parameters defining relationships between all the variables. Within the BNM each node in turn is considered a dependent variable (as in GLM/GLMM) and all other nodes as potential independent variables. A globally optimal DAG is then identified by a process of structure discovery or structure learning. There are different methods for identifying the best DAG depending on the number of variables present in the analysis. The exact search method based on a goodness of fit criterion i.e. the highest log marginal likelihood score (network score), was used in this study (Koivisto & Sood, 2004). The best DAG was identified by fitting 2^n models, where n is the number of variables (in this case 2^{20}= 1048576 models were fitted with each parent limit). Model complexity increases as the number of parents is increased.

Banning certain relationships

Certain relationships can be banned from selection based on *a priori* knowledge about their association. In this study, climate and land cover variables could influence host variables but the reverse relationships were banned from being selected in the network. This DAG was compared with the DAG obtained without any restrictions based on the network score, and the better of these two models (i.e. the one with the highest log likelihood score) was further used for bootstrapping.

Parametric bootstrapping

The best DAG identified in the exact search method can be considered as the final DAG, but there is always a chance of over fitting as with any other variable selection method (Babyak, 2004). This chance, however, is considered small for BNM, especially with the very large sample size available in the present study (more than 15,000 villages). Nevertheless, parametric bootstrapping was conducted to check overfitting and to delete arcs which are not important. Parametric bootstrapping was used to prune the arcs identified in the exact search method. This is a computationally intensive task that required cluster computing for its rapid implementation. In parametric bootstrapping, simulation is performed many times (50 or more times) to generate datasets (artificial) of similar size to the original (i.e. observed) dataset. Simulation to generate bootstrap datasets is performed by providing marginal posterior densities (initial values) for each and every parameter in the best DAG identified in the previous step. Simulation can be performed in any of the MCMC software; WinBugs or Just Another Gibbs Sampler, JAGS. The latter was used in the present study. The simulated bootstrapped datasets were again subjected to best DAG selection as described earlier. The main objective of bootstrapping is to check the percentage of arcs retained in >50% of the bootstrapped datasets (Friedman et. al 1999).

In JAGS, as mentioned above, we need to provide the structure of the dependencies, and the marginal posterior distribution of each parameter for simulating datasets, but actual data are not provided as we are using MCMC to simulate datasets for bootstrapping and perform BNM on the bootstrapped datasets. The number of MCMC iterations and thinning is decided based on the correlation between samples (autocorrelation plots) and it should give similar number of samples as in the original dataset (~ 16800 in this case). 300,000 iterations were performed with a 132,000 burn in and the actual MCMC is then run for 168,000 iterations with a thin of 10, which gives 16800 observations, the same sample size as the original data. Once the bootstrap datasets are generated, the exact search for identifying the best DAG performed on the original dataset needs to be repeated on all the bootstrapped datasets. The final step is to compare the number of times a particular arc is selected in bootstrap datasets which have been represented in 50% (or more) of the DAGs. This structure represents the relationship between variables and their dependencies (both direct and indirect). The Markov blanket is defined as that set of variables that is ether a parent or child of a particular variable under study. Here the variable of interest is the number of BTV cases so its Markov blanket includes only the parent and child nodes directly linked to it in the BNM and it was only these variables that were later used in modelling to estimate regression co-efficients or to make predictions in a GLM/GLMM framework. The BNM itself is not designed for making spatial predictions or extrapolating the results to other areas.

6.2.6 Bayesian geostatistical modelling approach

Relationships between bluetongue cases at the village level and environmental predictors (selected in BNM) were quantified using a Bayesian generalised linear mixed modelling approach with Poisson errors and by employing a geostatistical correlation structure to account for spatial autocorrelation. Suppose we have i=1,….n number of villages. Village locations (i) are spatially geo-referenced as S= {$s_{1,}$ …., s_N}; BTV spreads locally, leading to spatial autocorrelation of disease cases; and environmental variables that may be contributing to BTV severity are independently spatially auto correlated as well. Hence bluetongue cases are dependent not only on β predictor variables, but also on the weighted spatial neighbours. Therefore it is assumed that village-based BTV case numbers, Y(s_i), follow a Poisson distribution and that these case numbers are related to environmental variables (x), spatial processes (w), and village level independent spatial errors (ε) as shown in equation 2 below:

$$\Pr(\mathrm{Y} = \mathrm{y}/\mu) = Poisson(\mu_i) \tag{2}$$

$$\log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots\ldots + \beta_n x_{ni} + w(s_i) + \varepsilon(s_i)$$

$$w(s_i) \sim GP(0, C(S_1, S_2)) - \text{spatial process}$$

$$C(S_1, S_2) = \sigma^2 \rho(S_1, S_2; \phi)$$

$$\rho(S_1, S_2; \phi) = \exp\left(-\frac{S_1, S_2}{\phi}\right)$$

$$\phi - \text{spatial decay and smoothness parameter}$$

$$\varepsilon(s_i) \sim N(0, \tau^2) - \text{ independent process (nugget)}$$

$$\text{Priors for different parameters}$$

$$\beta \sim N(\mu_\beta, \Sigma_\beta) \text{ - regression co - efficients}$$

$$\sigma^2 \sim IG(a, b) - \text{variance}$$

$$\phi \sim U(c, d) - \text{correlation parameter (range)}$$

$$\tau^2 \sim IG(e, f) - \text{nugget variance}$$

The residual variance consists of a spatial process $w(s_i)$ and an independent process $\varepsilon(s_i)$. The independent process is modelled using the observed nugget variance as the prior. The spatial process or spatial random effects accounts for the spatial dependence caused by intrinsic processes such as disease spread or by unmeasured or unobserved covariates. The spatial process will not account for missing covariates which do not have spatial structure and this additional variance is accounted for by the independent process. The spatial process is specified by a zero mean Gaussian process and covariance function $C(S_1, S_2)$. The correlation function $\rho(S_1, S_2; \phi)$ varies with the correlation structure (Gaussian, Spherical, Exponential) specified and ϕ denotes the spatial decay and smoothness parameter. An exponential model was fitted to the present dataset with covariates. Prior distributions for the parameters are defined to complete the hierarchical structure of the model. The regression co-efficients β are assigned multivariate Gaussian priors. The variance component of the independent process (*nugget*) τ^2 and variance of the spatial effect σ^2 are assigned inverse-Gamma priors. The correlation parameter (spatial decay and smoothness) ϕ are assigned a uniform prior distribution. The range of this uniform prior is decided based on the residual semi-variogram to cover a broad range of spatial autocorrelation.

Equation (1) will be very difficult to implement when n is large and so a predictive process was employed to reduce the n by placing knots on the observations (location of BTV records) using two clustering algorithms (*k-means* and *k-medoids*) and equation (1) can be rewritten with only a subset of n being used. All the assumptions and priors remain the same.

Let n* be the number of knots with n*<<n, and w*= (w(s*$_{1)\ldots}$w (s*$_n$))

The projected spatial process w(s) at locations s, based on the knots can be given by a "kriging equation" w$^{\sim}$ (s) = E {w(s) |w*},

Where w*= (w(s*1), w(s*2),…w(s*n))

W$^{\sim}$(s) is known as the predictive process derived from the parent spatial process w(s)

Equation 2 can be re-written with the predictive process model

$$\Pr(\mathrm{Y} = \mathrm{y}/\mu) = Poisson(\mu_i) \qquad (3)$$

$$\log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_n x_{ni} + w^{\sim}(s_i) + \varepsilon(s_i)$$

The predictive process model (Equation 3) also accounts for non-stationarity (trend) in the data.

Model building

Bluetongue case data from Andhra Pradesh were divided into training and test data, 75% and 25% respectively. The relationships between the maximum number of bluetongue cases and the independent variables were quantified using a non-spatial Bayesian regression model. Semivariogram models were fitted on the residuals to check for spatial autocorrelation and to assess the best correlation structure to be specified in the Bayesian spatial model. The selection of a semi-variogram model can also be done in a Bayesian framework, but it can be computationally expensive to try all the correlation structures, particularly given a large sample size. The range and sill of the semi-variogram was used to decide the prior distribution for ϕ.

The next step was to create "knots" to reduce the number of observations (location of villages in Andhra Pradesh) for estimating the spatial correlation parameter. Two knot selection procedures were employed with knots at 64, 128, 256 and 512 knots. The approximate number of villages in each cluster with knots at 64, 128, 256 and 512 were 193, 96, 48, and 24 respectively. Models with two clustering algorithms and four different knots resulted in fitting of 8 models and the best knot model was selected based on reduced DIC (Spiegelhalter et al., 2002).

The best knot model (Bayesian spatial model with covariates) was used to make predictions on hold out test data and the predictions were compared with the observed bluetongue cases in Andhra Pradesh. The model predictions were compared with the observed bluetongue cases using Root Mean Square Errors (RMSE).

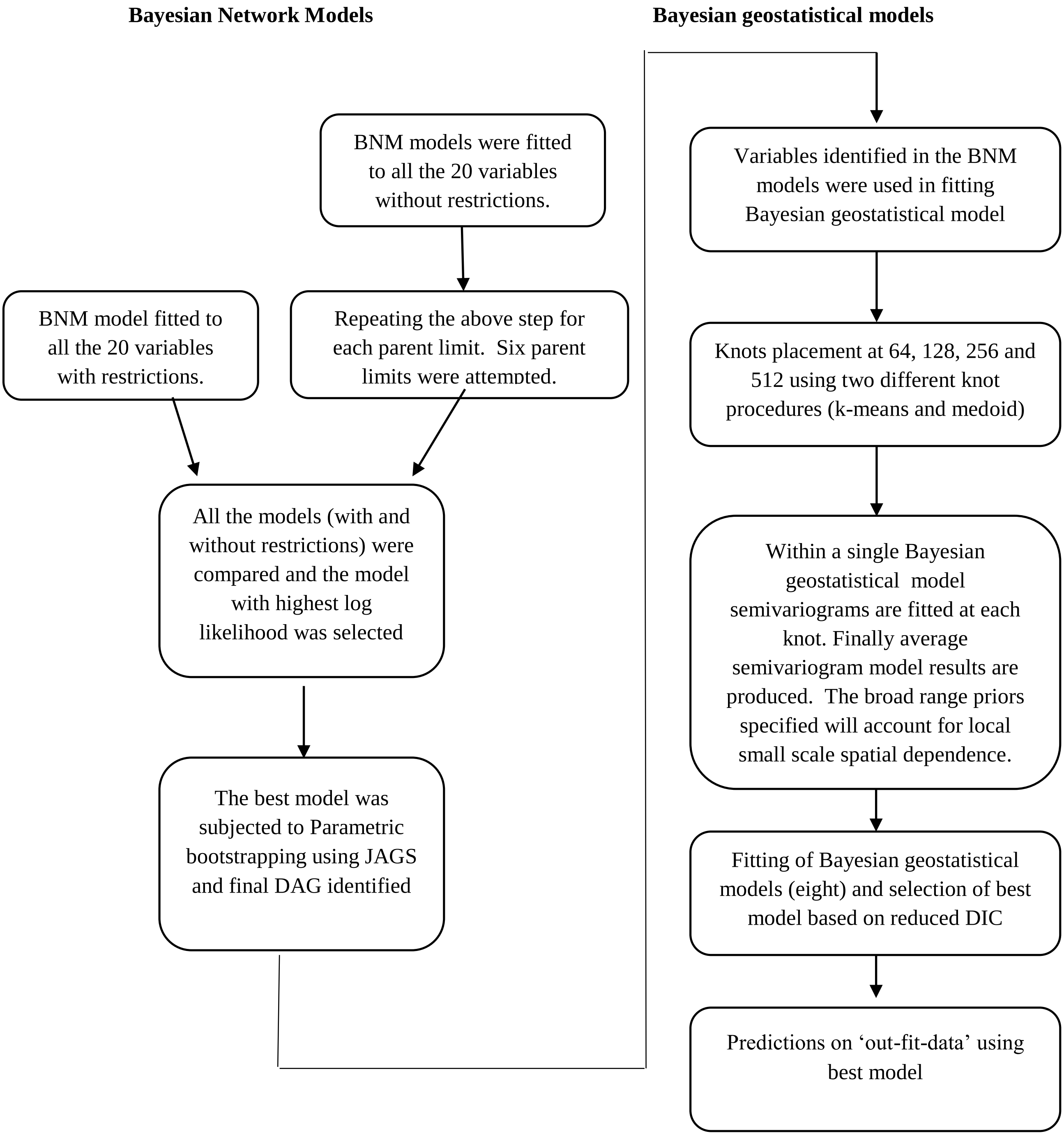

Bayesian Network Models
Bayesian geostatistical models
BNM models were fitted to all the 20 variables without restrictions.
BNM model fitted to all the 20 variables with restrictions.
Repeating the above step for each parent limit. Six parent limits were attempted.
All the models (with and without restrictions) were compared and the model with highest log likelihood was selected
The best model was subjected to Parametric bootstrapping using JAGS and final DAG identified
Variables identified in the BNM models were used in fitting Bayesian geostatistical model
Knots placement at 64, 128, 256 and 512 using two different knot procedures (k-means and medoid)
Within a single Bayesian geostatistical model semivariograms are fitted at each knot. Finally average semivariogram model results are produced. The broad range priors specified will account for local small scale spatial dependence.
Fitting of Bayesian geostatistical models (eight) and selection of best model based on reduced DIC
Predictions on 'out-fit-data' using best model

Figure 6.3: Flow chart explaining steps for Bayesian Network Modelling and Bayesian geostatistical method.

6.3 Results

6.3.1 Bayesian Network Modelling results

Out of more than 25000 villages in Andhra Pradesh, 16511 villages had sheep populations but only 524 villages ever reported BTV outbreaks during the study period (2004-2011). The maximum number of bluetongue cases (1800) over the study period was observed in Naguluppala padu village in the Prakasam district of Andhra Pradesh. Bluetongue outbreaks (Figure 6.7A) occur more in areas (South and North-West region) of high sheep population (Fig. 6.7B).

There are more indigenous cattle in Southern and Northern villages and fewer in some Central villages (Fig. 6.7A). Non-descript sheep populations are higher in some Southern areas and lower in the some central and North-Eastern part of Andhra Pradesh (Fig. 6.7B).

Central and North-Eastern villages are more irrigated (post-flooding or irrigated croplands) than other villages (Fig. 6.7C). There are more rain fed croplands in North-Eastern villages than elsewhere (Fig. 6.7D). The annual mean temperature is higher in Northern and Central villages of Andhra Pradesh and lower in Western and North-Eastern villages (Fig. 6.7E) which form part of the Western and Eastern Ghats respectively. These Ghats (both Eastern and Western) receive higher rainfall and are covered by more forest compared to other parts. The Northern, Eastern and coastal parts of Andhra Pradesh receive more South-West monsoon rainfall compared with the central and Southern parts (Fig. 6.7F).

The log marginal likelihood (marginal score) of BNM models when fitted by increasing the parent limit is shown in Table 6.2. The complexity increases as the parent limit increases and the BNM model with a six-parent limit is, perhaps unsurprisingly, best. Therefore, further analysis was restricted to a six-parent DAG without restrictions. The DAG with restrictions performed poorly compared to the unrestricted DAG with six parents, five parents, or four parents.

The six-parent unrestricted DAG is shown in Fig. 6.4 and after bootstrapping in Fig. 6.5. The best DAG after bootstrapping (Fig. 6.5) lost only two arcs of the original DAG without bootstrapping; the association of exotic & cross bred sheep with the Deccani breed and of terrain aspect with the land cover class ‘closed needle-leaved evergreen forest’. The bootstrapped DAG also lost the node of ‘exotic and crossbred sheep’ which was not associated with any of the other variables in the DAG. The importance and impacts of the links in the final DAG after bootstrapping are shown by their coefficients (Tables 6.3, 6.4, 6.5 & 6.6). The coefficients of a single DAG are given in different tables for each node. Each node representing a variable is equivalent to a single GLM model consisting of a dependent (parent) variable and independent variables (child nodes). Therefore each node should be interpreted with their respective parent nodes and their relationships with child nodes. The co-efficients of the DAG (with six parent limit and after bootstrapping) are split into different categories (BTV cases, host nodes, land cover nodes and climate nodes) (Tables 6.3, 6.4, 6.5 & 6.6).

Unrestricted DAG model results (Tables 6.3, 6.4, 6.5 & 6.6)

Impacts of climate, land use and hosts on bluetongue cases in Andhra Pradesh

Bluetongue cases increase with increases in annual mean temperature, densities of indigenous cattle and non-descript sheep and coverage of post-flooding or irrigated croplands, but decrease with increases in South-West monsoon rainfall, and coverage of rain fed croplands (Fig. 6.5).

How does climate constrain land use in Andhra Pradesh?

Rain-fed croplands, post-flooding or irrigated croplands, and closed broadleaved deciduous forest all increase in areas of high annual rainfall. However, relationships between rainfall and land use are more subtle and depend on monsoon season. Post-flooding or irrigated croplands are found in areas of high North East monsoon rainfall but low South West monsoon rainfall whilst rain fed croplands are found in areas with low South-West monsoon rainfall (Fig. 6.5).

How does climate and land use constrain host densities in Andhra Pradesh?

Non-descript sheep densities tend to be higher in areas of rain fed croplands, areas with high South West monsoon rainfall and lower in areas of closed broad leaved deciduous forest or areas with steep slopes. They are also highly positively correlated with buffalo densities.

Indigenous cattle densities also increase with the coverage of rain fed croplands and some forest types (closed to open broadleaved evergreen and/or semi-deciduous forest, closed needle leaved evergreen forest), but decrease on terrain with increasing gradients (slopes).

Densities of Deccani sheep decrease with increases in areas of closed broadleaved deciduous forest and in areas of increasing slope, whereas they are positively associated with Indigenous cattle and buffalo densities, rain fed croplands and South West monsoon rainfall. Non-descript sheep populations' increase with increases in buffalo population, rain fed croplands and South West monsoon rainfall and decrease with increases in Deccani sheep, closed broadleaved deciduous forest and terrain slope.

Rain fed croplands support high densities of buffalo and are positively associated with mosaic vegetation and slope. Post-flooding or irrigated cropland also supports high densities of buffalo and is positively associated with rain fed croplands, mosaic vegetation, close to open type of forest, closed forest and slope. Mosaic cropland supports high densities of indigenous cattle and is positively associated with rain fed croplands, post-flooding or irrigated croplands and mosaic vegetation, whereas it is negatively associated with closed to open forest (VALUE_40 in Table 6.1) and North-East monsoon rainfall.

Indigenous cattle and Buffalo densities increase with increases in temperature. Deccani sheep and non-descript sheep, however, both decrease with increases in temperature. Temperature is positively associated with slope and negatively associated with South West monsoon rainfall. North East monsoon rainfall supports high densities of indigenous cattle, but does not support buffalo, Deccani sheep or non-descript sheep.

Increases in South-West monsoon rainfall are associated with increases of indigenous cattle populations and decreases of closed broadleaved deciduous forest and terrain slope. Annual rainfall is negatively associated with buffalo, rain fed croplands and post-flooding or irrigated cropland and negatively associated with mosaic vegetation.

Relationships among the land cover variables and climate

Mosaic vegetation is positively associated with closed type of forest and negatively associated with closed to open type of forest. Closed to open broadleaved and/or semi-deciduous forest is positively associated with closed type of forest. Closed needle leaved evergreen forest is positively associated with mosaic vegetation, closed to open the type of forest, closed broadleaved deciduous forest and mosaic forest. Aspect is positively associated with slope, whereas it is negatively associated with temperature.

6.3.2 Restricted DAG model results (Fig 6.6)

The unrestricted DAG identified some of the relationships which may not be plausible. The spurious relationships identified include climate variables (temperature and rainfall) dependent on the host (sheep, buffalo and cattle) and land cover (irrigated and rain fed croplands) influenced by densities of buffalo. The spurious relationships were not identified in the DAG (Fig. 6.6) when certain arcs were banned by assuming certain restrictions like the host influencing climate, land cover or aspect and slope.

The variables directly associated with bluetongue cases in the restricted DAG model were slightly different from the variables identified in the unrestricted DAG. The annual mean temperature, densities of indigenous cattle, South-West monsoon rainfall, post-flooding or irrigated croplands were also identified in the restricted DAG (Fig. 6.6). The mosaic cropland and closed broad deciduous forest was identified in the restricted model.

6.3.3 Spatial regression (Bayesian geostatistical) model results

The residuals of the non-spatial model relating BTV outbreaks indicate the presence of spatial autocorrelation (Fig.6.9). The minimum distance between two village centroids was 130 metres and the maximum was 888 km. Observed village locations and the locations of 256 knots placed by two different methods are shown in Fig. 6.8.

Comparison of DIC using the *k-means* clustering algorithm shows that the model with 128 knots is the most parsimonious model with lowest DIC (Table 6.7). The best model in the k-mediod based clustering is the one with 256 knots. Comparison of the two results shows that the model with 256 knots and medoid based clustering is better than other models.

Bluetongue cases are significantly positively associated with annual mean temperature, indigenous cattle numbers, non-descript sheep numbers and post-flooding or irrigated croplands and significantly negatively associated with rain fed croplands and South-West monsoon rainfall. The relationships and the signs of the co-efficients are similar to the final bootstrapped DAG (Fig. 6.5), but the magnitude of each effect changes considerably in the spatial regression model with incorporation of spatial autocorrelation (Table 6.8). The spatial decay parameter ϕ is estimated as 175 km (95% credible interval; 31km-434km).

Comparing the fit of the model (Fig 6.11) shows that predictions coincide quite well spatially with observations but overall the model predicted more cases than were observed, and the overall correlation between observed and predicted outbreaks is poor (r=0.21). The Root Mean Square (RMSE) statistics for the model on training and test data are given in Table 6.9.

Variable name	Description	Variable type
BT_cases	Bluetongue affected maximum number of cases	Poisson
Ind_Cattle	Indigenous cattle	Gaussian
Buffalo	Buffalo	Gaussian
Exotic & cross_sheep	Exotic and crossbred sheep	Gaussian
Deccani	Deccani breed of sheep	Gaussian
ND_sheep	Non-descript sheep	Gaussian
VALUE_11	Post-flooding or irrigated croplands	Gaussian
VALUE_14	Rain fed croplands	Gaussian
VALUE_20	Mosaic cropland (50-70%)/vegetation (grassland, shrub land, and forest) (20-50%)	Gaussian
VALUE_30	Mosaic vegetation (grassland, shrub land, forest) (50-70%)/cropland (20-50%)	Gaussian
VALUE_40	Closed to open (>15%) broadleaved evergreen and/or semi deciduous forest (>5m)	Gaussian
VALUE_50	Closed (>40%) broadleaved deciduous forest (>5m)	Gaussian
VALUE_70	Closed (>40%) needle leaved evergreen forest (>5m)	Gaussian
VALUE_110	Mosaic forest/shrub land (50-70%)/ grassland (20-50%)	Gaussian
Temperature	Annual mean temperature	Gaussian
Aspect	Aspect	Gaussian
Slope	Slope	Gaussian
Annual_rain	Annual monsoon rainfall	Gaussian
NE_rain	North east monsoon rainfall	Gaussian
SW_rain	South west monsoon rainfall	Gaussian

Table 6.1: Variables used in the Bayesian network analysis. The variable labels used in the DAGs (Figures 6.4, 6.5 and 6.6) are shown in the first column.

Parent limit	**Log marginal likelihood (network score)**
Six parents limit with no restrictions	**-507752**
Five parents limit with no restrictions	-508210
Four parents limit with no restrictions	-509382
Banned arcs with six parents limit (with restrictions)	-510128
Three parents limit with no restrictions	-511414
Two parents limit with no restrictions	-513923
One parent limit with no restrictions	-519605

Table 6.2: Log marginal likelihood (network score) of DAG's with different parent limits and with banned relationships. The higher the network score, the better the model.

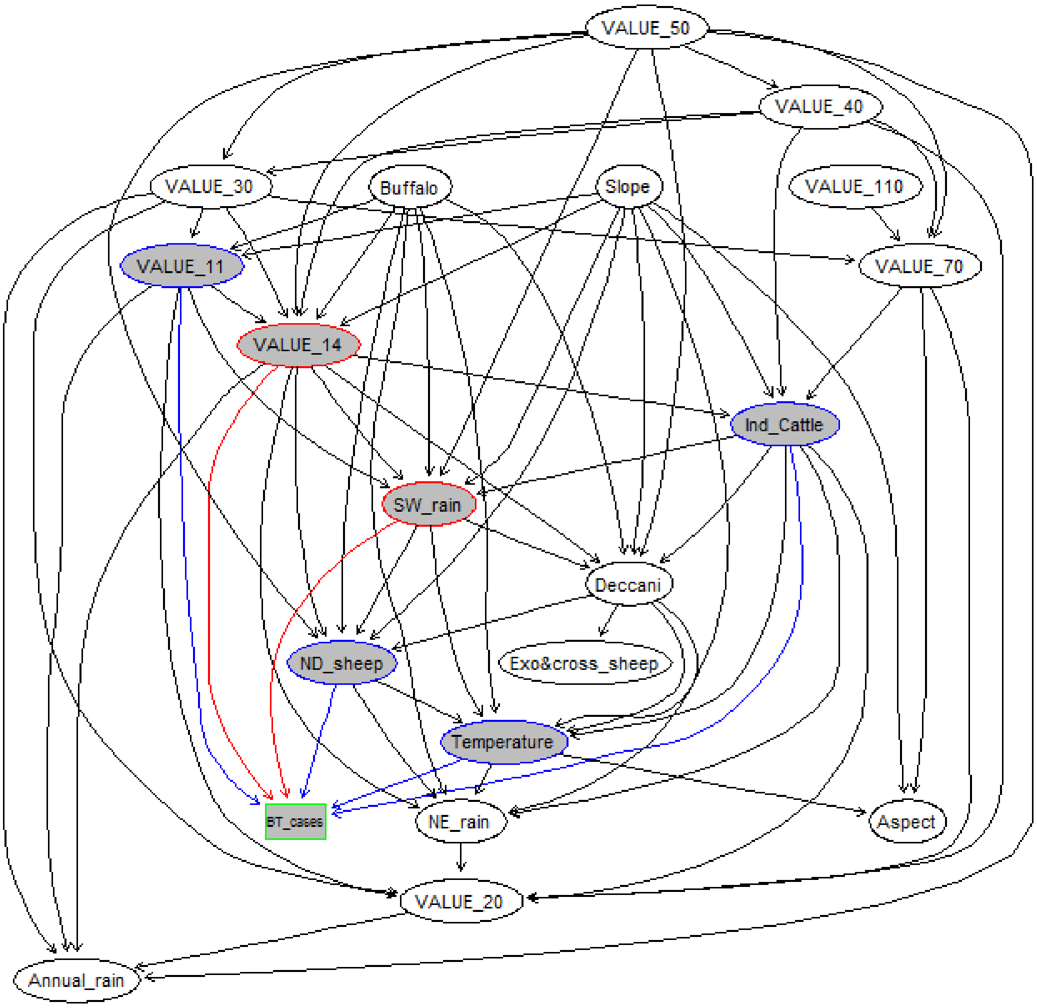


Figure 6.4: The final DAG with a six parent limit (unrestricted).Normally distributed variables are shown in ellipses and Poisson-distributed bluetongue cases are shown in a box. The information flows from parent to child and the direction of the arrows shows the dependency of the variables with other variables. The blue arrows towards BT_cases show positive associations and red arrows show negative associations. The filled ellipses (grey color) show the variables which are directly related to bluetongue cases. See Table 1 for the full names of the variables.

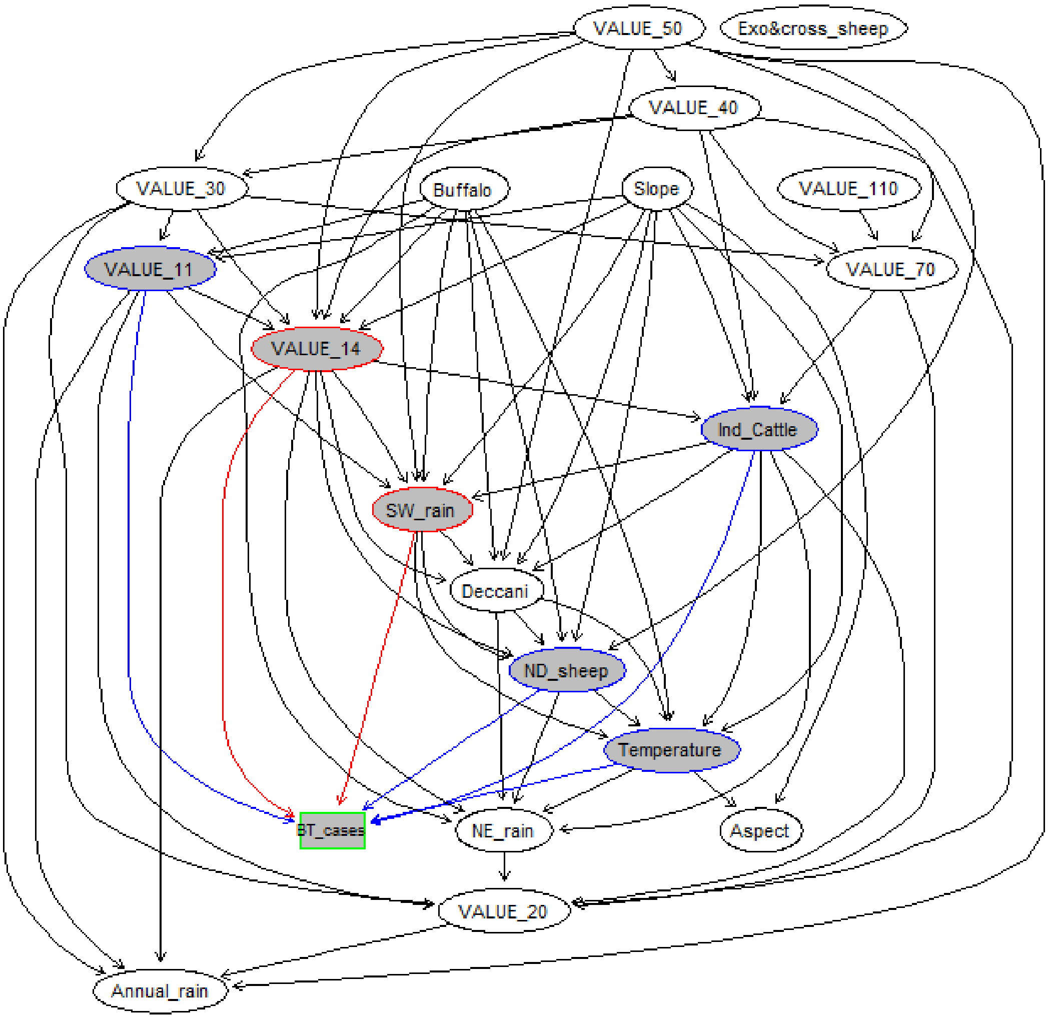


Figure 6.5: The final DAG with six parent limit (unrestricted) after bootstrapping (>50% retained in the bootstrapping).Normally distributed variables are shown in ellipses and Poisson-distributed bluetongue cases are shown in a box. The information flows from parent to child and the direction of the arrow shows the dependency of the variable with particular variable. The blue arrows towards BT_cases show positive association and red arrows show negative association. The filled ellipses (grey color) show the variables which are directly related to bluetongue cases. The exotic and crossbred sheep variable is not related to any of the variables in this DAG. See Table 1 for the full names of the variables.

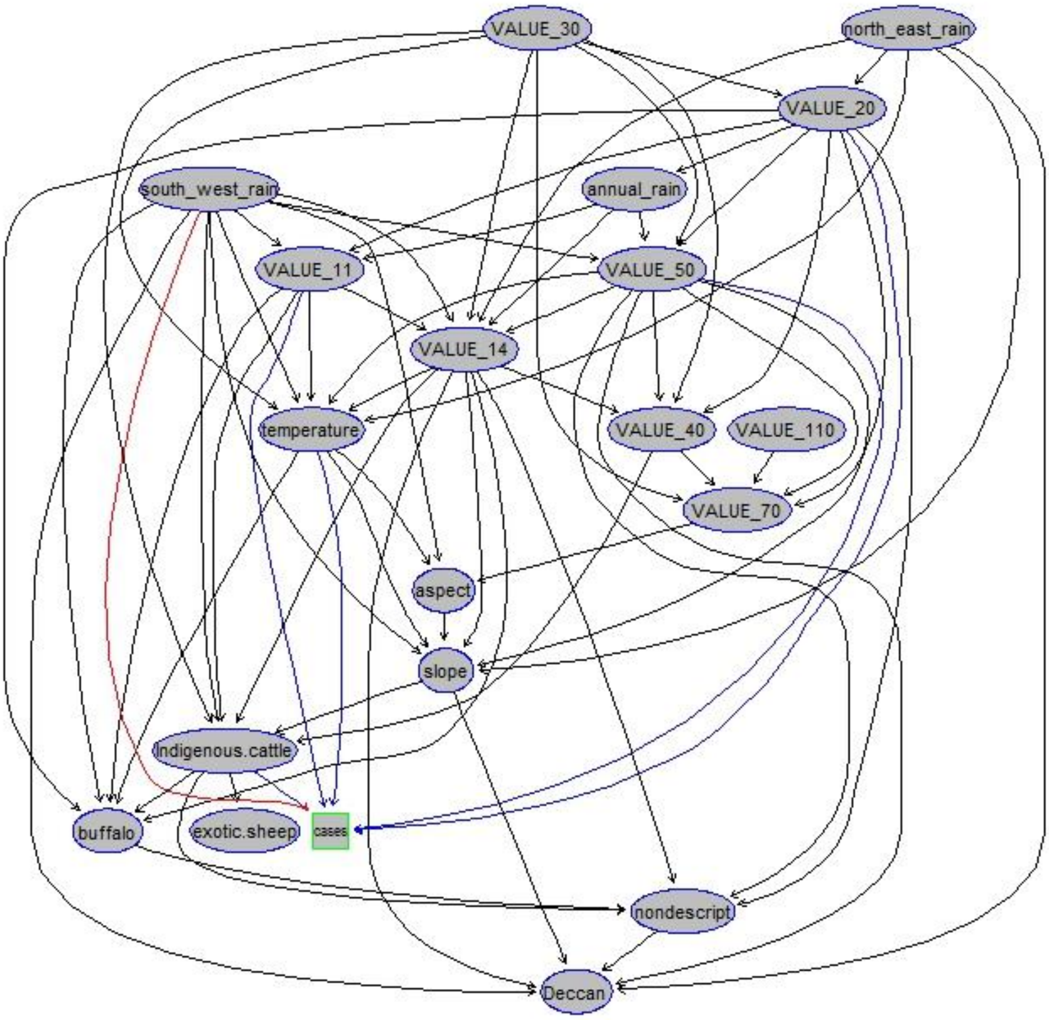


Figure 6.6: The DAG (restricted or banned model) with six parent limit. Normally distributed variables are shown in ellipses and Poisson-distributed bluetongue cases are shown in a box. The information flows from parent to child and the direction of the arrow shows the dependency of the variable with particular variable. The blue arrows towards BT_cases show positive association and red arrows show negative association. The filled ellipses (grey color) show the variables which are directly related to bluetongue cases. See Table 1 for the full names of the variables.

Bluetongue cases	**Credible interval**
Indigenous Cattle	0.393, 0.414
Nondescript Sheep	0.268, 0.287
VALUE_11	0.261, 0.270
VALUE_14	-0.203, -0.185
Temperature	1.342, 1.430
South west monsoon rainfall	-0.418, -0.400

Table 6.3: Credible intervals of variables associated with the bluetongue node in Figure 5 (with a 6 parent limit and unrestricted DAG).Co-efficients in blue indicate a positive association and in red a negative association with bluetongue cases. All the relationships identified in the final DAG are significant because the credible intervals do not bridge zero.

Host Nodes	**Credible interval (95%)**
Indigenous Cattle	
VALUE_14	0.153, 0.175
VALUE_40	-0.074, -0.052
VALUE_70	-0.079, -0.058
Slope	-0.058, -0.038
(**Buffalo**	**No parent nodes**)
(**Exotic and crossbred sheep**	**No parent or child nodes**)
Deccani sheep	
Indigenous cattle	0.042, 0.062
Buffalo	0.065, 0.085
VALUE_14	0.057, 0.080
VALUE_50	-0.070, -0.048
Slope	-0.334, -0.312
South west monsoon rainfall	0.215, 0.238
Non-descript sheep	
Buffalo	0.107, 0.125
Deccani breed of sheep	-0.361, -0.342
VALUE_14	0.214, 0.235
VALUE_50	-0.119, -0.098
Slope	-0.186, -0.165
South west monsoon rainfall	0.384, 0.404

Table 6.4: Credible intervals of variables associated with host nodes (in bold) identified in the best DAG with a six parent limit (unrestricted).The co-efficients depicted in blue indicate a positive association and in red a negative association. All the relationships identified in the final DAG are significant because the credible intervals do not bridge zero.

land cover Nodes	**Credible interval (95%)**
Value_11	
Buffalo	0.250, 0.268
VALUE_30	0.276, 0.295
Slope	0.074, 0.093
Value_14	
Buffalo	0.148, 0.166
VALUE_11	0.143, 0.162
VALUE_30	0.182, 0.203
VALUE_40	0.138, 0.163
VALUE_50	0.211, 0.239
Slope	0.067, 0.085
VALUE_20	
Indigenous cattle	0.045, 0.055
VALUE_11	0.108, 0.119
VALUE_30	0.876, 0.889
VALUE_40	-0.051, -0.040
VALUE_70	-0.100, -0.088
North East monsoon rainfall	-0.027, -0.017
VALUE_30	
VALUE_40	-0.331, -0.305
VALUE_50	0.671, 0.697
VALUE_40	
VALUE_50	0.719, 0.733
(VALUE_50	**No parent nodes**)
VALUE_70	
VALUE_30	0.497, 0.516
VALUE_40	0.092, 0.117
VALUE_50	0.102, 0.129
VALUE_110	0.064, 0.080

Table 6.5: Credible intervals of variables associated with land cover nodes (in bold) identified in the best DAG with a six parent limit (unrestricted model).The co-efficients depicted in blue indicate a positive association and in red a negative association.

Climate Nodes	**Credible interval (95%)**
Temperature	
Indigenous cattle	0.019, 0.032
Buffalo	0.163, 0.175
Deccani sheep	-0.053, -0.039
Nondescript Sheep	-0.052, -0.039
Slope	0.803, 0.818
South west monsoon rainfall	-0.0636, -0.048
Aspect	
VALUE_70	0.022, 0.039
Temperature	-0.092, -0.065
Slope	0.631, 0.657
(Slope	**No parent nodes**)
Annual monsoon rainfall	
VALUE_11	0.110, 0.131
VALUE_14	0.165, 0.188
VALUE_20	0.107, 0.145
VALUE_30	-0.137, -0.098
VALUE_50	0.071, 0.094
North east monsoon rainfall	
Indigenous Cattle	0.125, 0.144
Buffalo	-0.114, -0.094
Deccani sheep	-0.164, -0.144
Nondescript Sheep	-0.293, -0.273
VALUE_14	-0.094, -0.074
Temperature	0.169, 0.189
South west monsoon rainfall	
Indigenous Cattle	0.241, 0.258
Buffalo	-0.095, -0.077
VALUE_11	-0.103, -0.085
VALUE_14	-0.137, -0.117
VALUE_50	0.113, 0.132
Slope	0.470, 0.487

Table 6.6: Credible intervals of variables associated with the terrain or climate nodes (in bold) identified in the best DAG with a six parent limit. The co-efficients depicted in blue indicate a positive association and in red a negative association. All the relationships identified in the final DAG are significant because the credible intervals do not bridge zero.

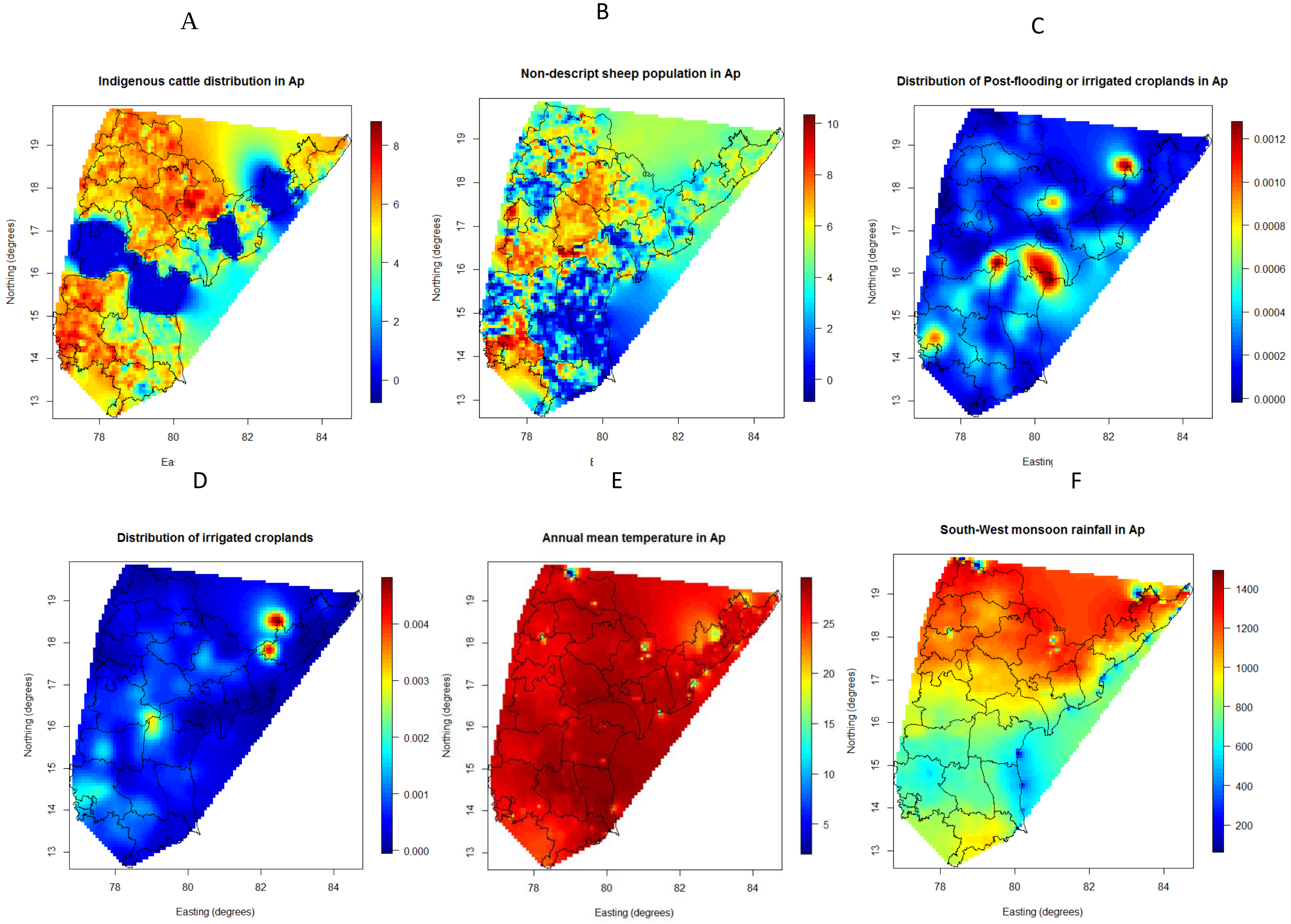

A
B
C
Indigenous cattle distribution in Ap
Non-descript sheep population in Ap
Distribution of Post-flooding or irrigated croplands in Ap
Northing (degrees)
13
14
15
16
17
18
19
78
80
82
84
Ea
E
Easting
0
2
4
6
8
10
0.0000
0.0002
0.0004
0.0006
0.0008
0.0010
0.0012
D
E
F
Distribution of irrigated croplands
Annual mean temperature in Ap
South-West monsoon rainfall in Ap
Easting (degrees)
0.000
0.001
0.002
0.003
0.004
5
10
15
20
25
200
400
600
800
1000
1200
1400

Figure 6.7: Distribution of different independent variables in Andhra Pradesh (AP) used in the analysis (A) logged (natural logarithm) Indigenous cattle (B)logged (natural logarithm) non-descript sheep population (C) area covered by post-flooding or irrigated croplands (D) area covered by rain fed croplands (E) Annual mean temperature (^{0}C) and (F) South-West monsoon rainfall (mm).

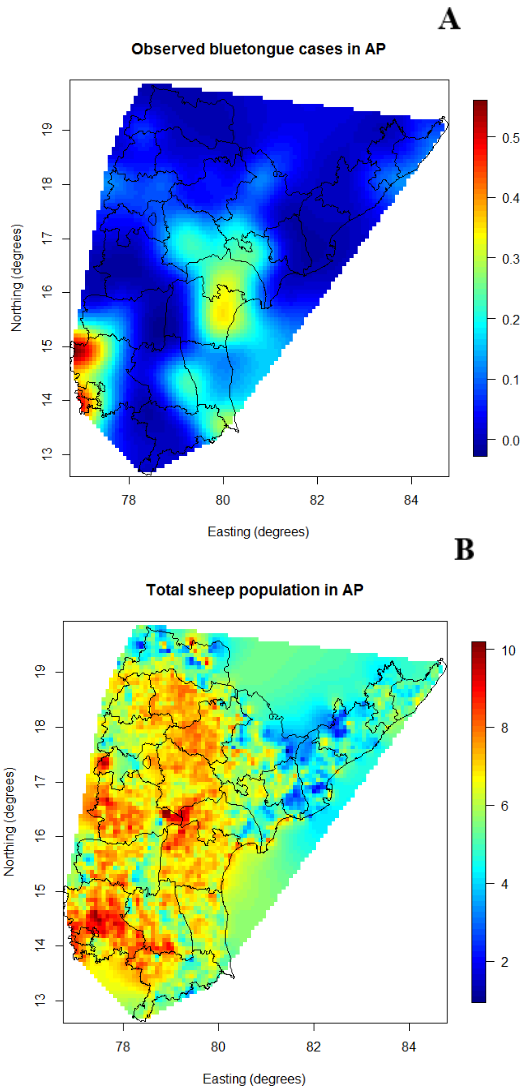


Figure 6.8: (A) Observed maximum bluetongue cases (natural logarithm) and (B) Total sheep population (natural logarithm) in Andhra Pradesh.

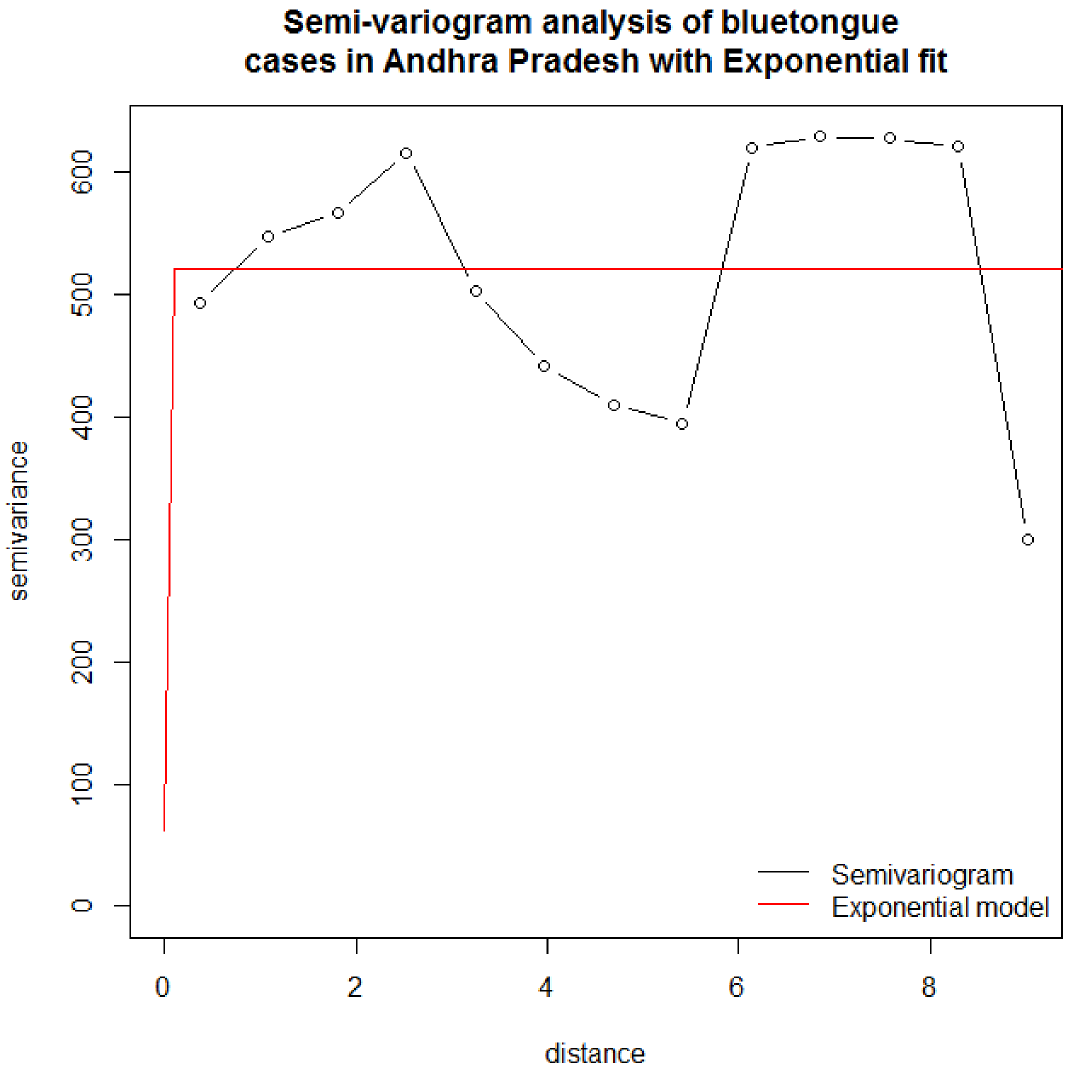


Figure 6.9: Semivariogram on the residuals of the non-spatial model. Dots indicates the empirical semi-variogram. The x-axis is distance in km and y-axis indicates the semi-variance.

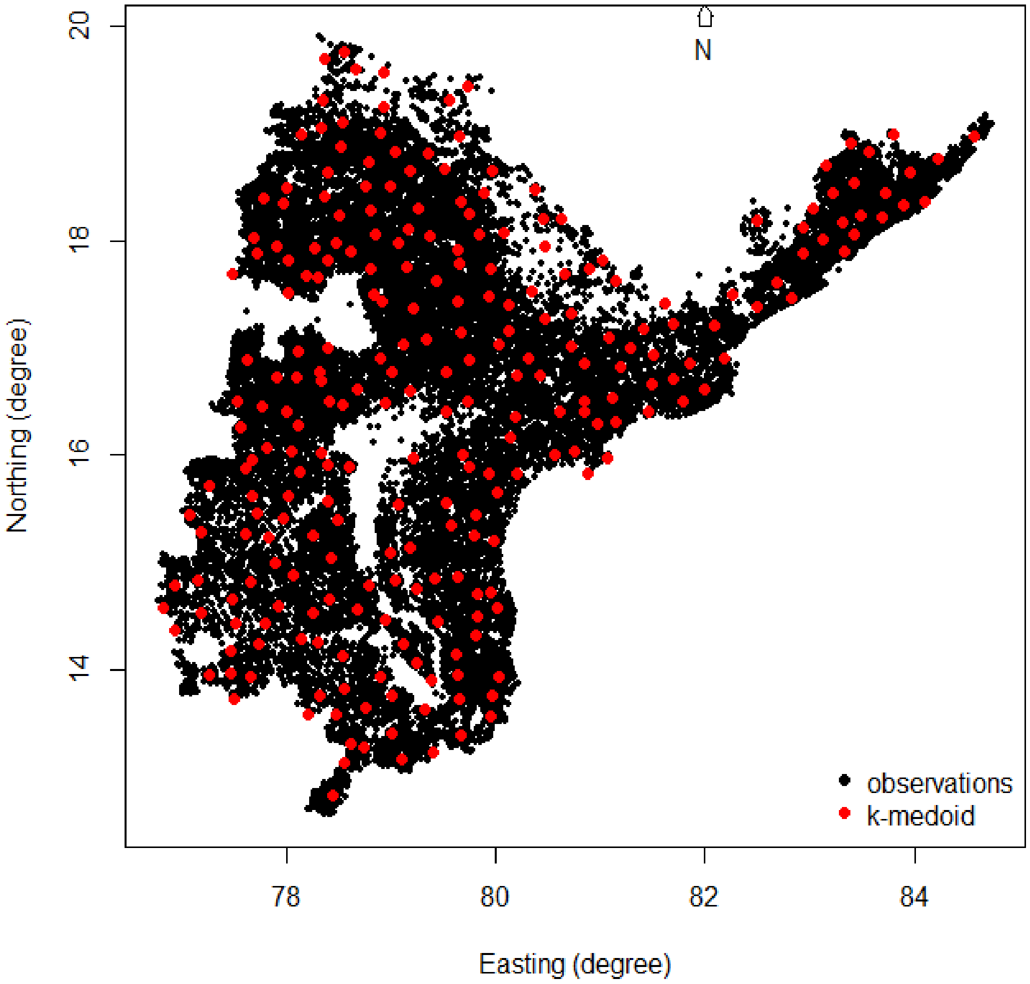


Figure 6.10: Plot of the locations of centroids used to cluster the villages of AP for measurement of semi-variance and spatial structure. The black dots show all the villages of Andhra Pradesh. The k medoid knots (n = 256) are shown as red dots. In medoid clustering the knot must be located at one of the actual observations. (Co-ordinate reference system: WGS 1984)

Model	64 knots	128 knots	256 knots	512 knots
Centroid based clustering	51229.18	**28254.33**	29764.1	49760.18
Medoid based clustering	36860.21	38960.13	**17653.33**	50096.57

Table 6.7: DIC of different models fitted using two clustering methods to generate 4 different knots in the observations. In medoid clustering the knot must be located at one of the actual observations.

Variable	**Mean**	**Credible interval**
Intercept	-7.891	-10.04, -6.67
Indigenous cattle	1.290	0.10, 1.41
Non-descript sheep	6.272	0.05, 7.36
Post-flooding or irrigated croplands	1.020	0.89 , 1.10
Rainfed croplands	**-6.462**	**-7.13, -4.70**
Mean temperature	3.136	0.027, 3.89
South-West Monsoon rainfall	**-1.906**	**-0.002, -1.54**
Sigma square	7.983	1.22, 4.74
Phi	1.583 (175km)	0.28, 3.91 (31-434 km)

Table 6.8: Mean coefficients and their credible intervals of the Bayesian spatial regression model (Equation 1) with 256 knots using the k-medoids based algorithm.

Variance of training data	RMSE (training data)	Variance of test data	RMSE (test data)
285.69	16.61	326.36	18.87

Table 6.9: Root mean square error (RMSE) statistics for the training and test data for the Bayesian spatial regression model with 256-knots-k-medoids.

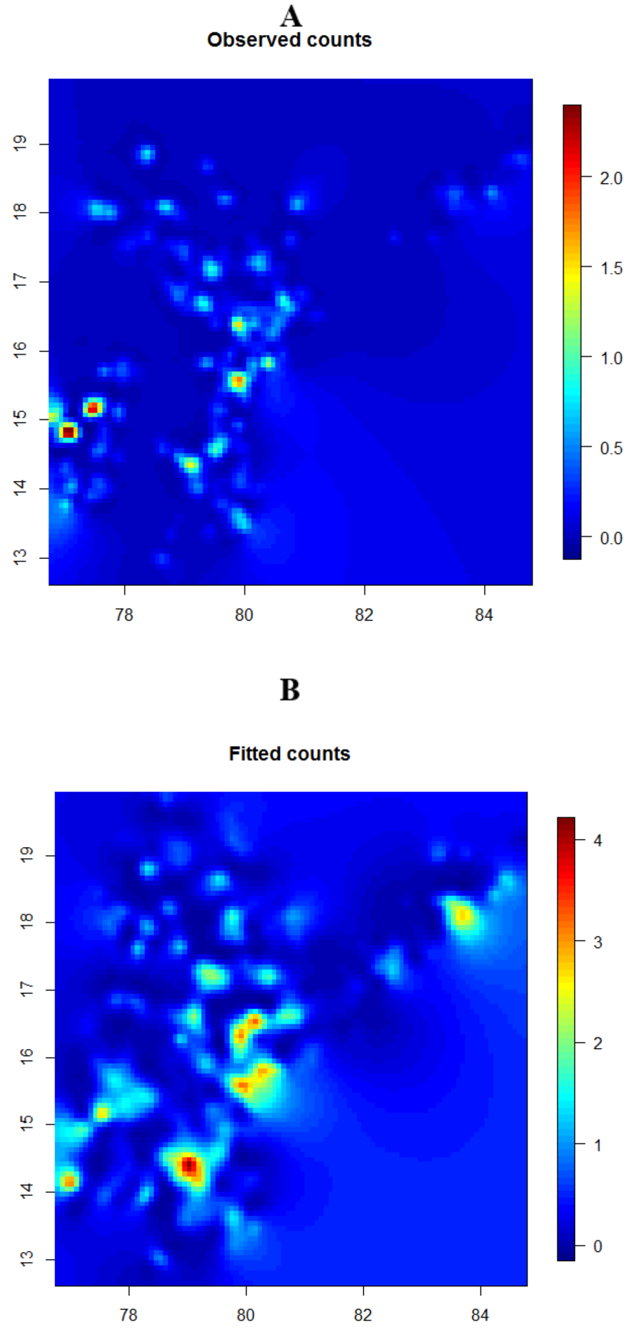


Figure 6.11: Observed (A) and fitted (B) bluetongue outbreaks using Bayesian spatial regression model (Equation 1) with 256 knots using the k-medoids based algorithm.

6.4 Discussion

The BNM identified associations of bluetongue with indigenous cattle, non-descript sheep, post-flooding or irrigated croplands and temperature (all positively associated) and rain fed croplands and South-West monsoon rainfall (both negatively associated) with bluetongue cases. These results cannot be directly compared with the variables identified in chapter 4 because the spatial scale is different and also the measure of bluetongue (mean number of BTV outbreaks in chapter 4 and maximum number of BTV cases in this chapter). The problem by aggregation of data and varying results at different spatial scales is referred to as the Modifiable Area Unit Problem (MAU) (Lawson, 2013).

Positive association of bluetongue cases with indigenous cattle can be attributed to the mixed farming system practiced in Andhra Pradesh by small and marginal farmers. Indigenous cattle are maintained for draft purpose and have low milk yields (Rao et al., 2010). The buffalo population increased after the green revolution in India, but still it is less than the cattle population. Every ten-fold increase in Indigenous cattle (i.e. a unit increase on the log. scale to which the data were transformed) increases the maximum BTV outbreaks by 1.29. The role of non-descript sheep is more important; for every ten-fold increase in these livestock, the maximum BTV outbreaks increased by more than 6.27 (Table 6.8).

Selection of post-flooding or irrigated croplands is supported by the fact that Andhra Pradesh contributes 13% of the rice produced in India (Adusumilli & Laxmi, 2011) and rice cultivation is dependent on irrigation systems. The larvae of the midges *C.oxystoma* and *C.arakawae* along with larvae of six other *Culicoides* species were found in active in abandoned rice fields (Yanase et al., 2013).

Temperature not only speeds up the extrinsic incubation period (EIP) but also has a significant effect on development and mortality of the potential vectors (Gubbins et al., 2008).

In the present BNM analysis buffalo numbers were not directly linked to BTV case numbers, but only indirectly related. Instead buffalo appear in a separate network that includes Deccani sheep, non-descript sheep, rain fed croplands, post-flooding or irrigated croplands and temperature (all positively associated) and the North-East monsoon rainfall and South-West monsoon rainfall (both negatively associated).

Annual rainfall, North-East monsoon rainfall, Deccani sheep and mosaic cropland were indirectly associated with bluetongue. Annual rainfall is directly related to post-flooding or irrigated croplands and rain fed croplands and this relationship was as expected. This relationship might not have been detected with other methods due to multicollinearity. Multicollinearity may not only be due to two variables related to each other, but also due to linear combinations of more than two predictor variables correlated with other variables (Dohoo et al., 1997) leading to unstable regression co-efficients and inflated estimates of standard errors. The BNM approach accounts for this multicollinearity by joint modelling of all the variables and identifying relationships which are determining bluetongue cases and also their inter-relationships (Fig. 6.5).

Similarly North-East monsoon rainfall is directly associated with temperature. The positive relationship of North-East monsoon rainfall with temperature is significant as the coastal regions and Southern Andhra Pradesh receive most of the North-East monsoon rainfall and these are also regions with high temperatures. Deccani sheep numbers are correlated with South-West monsoon rainfall, indigenous cattle and rain fed croplands. As expected the temperature and slope were directly related to each other. Climate variables were also found to be co-dependent in studying the association of weather factors with different pig pathologies (McCormick et al., 2013).

As expected, most of the land cover variables were related with each other each other and annual rainfall is directly associated with five of the eight land cover variables.

The problem of multicollinearity can be accounted for by exploratory correlation analysis and excluding highly correlated variables, but the criteria (level of correlation co-efficient) for selecting or deleting a variable is arbitrary (Dohoo et al., 1997). Multicollinearity will be a problem when the correlation co-efficient between two variables is >0.9, but it can also cause problems at lower levels depending on the variables under study. The level of correlation is difficult to identify when there is linear correlation of the predictor variables with other variables and this association can be identified in the Bayesian network modelling.

Spatial regression analysis discussion

The aim of the Bayesian spatial regression was to quantify the role of different predictor variables (identified in the BNM approach) on bluetongue cases in Andhra Pradesh and to check whether the inclusion of spatial autocorrelation changes the direction and significance of the variables related to BTV cases and involved different methods to handle the large spatial dataset. The Bayesian hierarchical model with spatial structure in the random effects allowed for accurate estimation of the parameters and their associated uncertainty.

Comparison of predictive process models with varying number of knots resulted in a better model with 128 knots in *k-means* method of clustering to group the observation for better estimation of correlation structure than 128 knots-k-medoid, but medoids based clustering method to generate 256 knots outperformed the 256 knot- *k-means* method. The medoid based clustering method is considered to be superior over *k-means* (Banerjee & Fuentes, 2012) because the knots are constrained to be from the observed locations.

The spatial predictive process based models (knots) will perform poorly when there is fine scale spatial range is less than the range specified in the knots (Finley et al., 2009) and closer knots are required to capture the small-scale spatial dependence. However, in this study the spatial range was approximately 175km (Table 6.8), so the knots based models can capture the spatial dependence in the data within the 95% credible interval of the spatial range (31-434 km). The range of spatial autocorrelation estimated in the spatial regression model can be due to some missing covariates or due to movement of animals (migration), which is very common practice in Andhra Pradesh. In a recent study on different aspects of sheep farming in Andhra Pradesh (Rao et al., 2013) found that the minimum and maximum distance migrated by sheep flocks was 51 km and 199 km respectively.

The significance of the variables and their associations (positive and negative) does not change when the BNM and spatial regression model results are compared, but the magnitude of the co-efficients do change. These changes are expected because the spatial regression model also accounts for spatial autocorrelation using the knots approach. Overall, the BNM identified important variables associated with bluetongue cases in Andhra Pradesh. The BNM also identified inter-dependencies between other variables (host, climate and land cover) which can be misleading sometimes and should be interpreted based on the biological understanding of the system. The possibility of including all the variables (hidden) is rare in any study and therefore can lead to false dependencies. The advantage of BNM in restricting certain relationships (if known *a priori*) is helpful in identifying the true dependencies. In this study, an attempt was made to include as many variables (climate, host, land cover) as possible to identify the true dependencies at a very high resolution (village level) for whole of Andhra Pradesh with a huge sample size. The spurious relationships identified in the present study (host influencing climate or land cover) may be due to non-inclusion of certain hidden variables (for example human population at village level). The use of Bayesian geostatistical model resulted in a poor fit to the observed bluetongue cases and needs further improvement. However, the spatial methods described in this chapter can be used in combination with BNM to account for spatial autocorrelation (if any) and also to make predictions.

Chapter 7

Discussion and Conclusions

7.1 Introduction

Analyses of bluetongue outbreaks and cases using statistical models were conducted in this study to understand the risk factors determining the extent and severity of the disease and to develop predictive models at different spatial and temporal scales. Past studies on the epidemiology of bluetongue in India have modelled only the presence and absence of bluetongue at district level, using a spatial logistic regression techniques and without reporting the important risk factors. With a view to develop predictive models to help control this economically important disease, the thesis explores bluetongue epidemiology in South India in novel ways with the following broad research questions in mind:

1. Is the epidemiological system for bluetongue in South India the same in the various states, or is there any evidence for different systems in different areas?
2. What is the role of abiotic extrinsic factors (climate) in determining seasonal variability of BTV outbreaks?
3. Is there any evidence that biotic intrinsic factors (host, breeds) are important in determining the occurrence and severity of bluetongue outbreaks in South India?
4. How can a very large number of potential predictor variables be reduced to a manageable number before building models of BTV transmission in India?

5. Can bluetongue be adequately forecast (with low RMSE and high correlation between observed and predicted BTV) at different spatial and temporal scales?

7.2 Different epidemiological systems in South India

Bluetongue is endemic in South India with outbreaks occurring every year. It was not known to the researchers whether the epidemiology of bluetongue is different in each state or the same. Considering the diverse habitat requirements for potential vectors of bluetongue and the different serotypes reported from each state, there is a possibility of different epidemiological systems in each state. The NLDA approach (Chapter-2) resulted in selection of a model (high sensitivity and specificity) with different presence and absence groups (three in each group) and this can be attributed to the different sets of environmental conditions prevailing in South India. Similarly, spatial analysis of BTV outbreaks using individual state models at district level identified different variables in each state (Chapter 4). The presence of different breeds in each state a subset of which were selected alongside different land cover and climate variables in state-specific spatial models further suggests the existence of different epidemiological systems in South India. This is significant for future vector or virus surveillance (high risk areas) in different zones and ultimately for designing control strategies (vaccination or vector control). Therefore, future studies in the rest of India should focus on different agro-ecological zones to understand the reasons for the absence of clinical disease in the presence of sero-positivity.

7.3 Intrinsic and extrinsic factors in determining bluetongue variability

7.3.1 Intrinsic and extrinsic factors determining temporal variability of BTV outbreaks at state and district level

Teasing apart the role of intrinsic and extrinsic factors is not straightforward but can be undertaken using statistical models and other methods (wavelet analysis). Poisson models are the most commonly employed statistical methods for analysing count data in infectious disease epidemiology (Hii et al., 2009; Hii, Rocklöv, et al., 2012; Hii, Zhu, et al., 2012). Over-dispersion in count data can arise due to many factors. One of the main reasons is the presence of monthly variability in the number of outbreaks and this can be accounted for by using autoregressive errors or using Quasi Poisson methods (Hii et al., 2009). The use of AR (1) models using Bayesian methods can account both for overdispersion in the data and for missing variables with temporal structure. The missing variables with temporal structure can be due to the waxing and waning of immunity or to any other factors with temporal structure (e.g. unmeasured seasonal variables) as discussed in Chapter 3. Therefore, detection of significant and dominant autocorrelation in the Poisson model can be due to intrinsic factors (like herd immunity) or extrinsic factors (like seasonal climate variables). In the time series analysis of bluetongue outbreaks the addition of the AR(1) term in the Poisson model resulted in considerable reduction in DIC compared to the model with only meteorological variables. Although the autoregressive process is dominant (lower DIC than model with meteorological variables only) in the temporal variation of bluetongue outbreaks, addition of meteorological variables (extrinsic process) marginally improved the overall fit of the model. Future studies should include immunity data and/or serotype information and other seasonal variables for disentangling the role of intrinsic and extrinsic factors. The inclusion of harmonics of particular periodicity (six months, one year, two years and three years) mimicking the influence of immunity can be tested to rule out the role of intrinsic mechanisms (herd immunity) in determining the seasonality and inter-annual variability in BTV outbreaks. In the preliminary analysis (results not shown) to evaluate the relative role of monthly meteorological conditions versus harmonic cycles of lengths 12, 24, 36, and 48 months, only the 12 month cycle, but none of the longer cycles, was significant.

Only annual periodicity in the bluetongue outbreaks and rainfall series could be detected in the data. It is therefore expected that the extent of seasonal BTV outbreaks will in part be determined and/or limited by host immunity which will therefore tend to reduce the strength of the correlations between climate variables and BTV outbreaks. Therefore, the role of intrinsic and extrinsic factors is important in the temporal variability of BTV outbreaks in Andhra Pradesh.

Inter-annual variability in the BTV outbreaks in Andhra Pradesh can be due to long term changes in climate as shown by the presence of significant around two year periodicity (wavelet coherence analysis with rainfall). The absence of significant periodicities of more than one year in BTV outbreaks (and in the rainfall data) data and detection of sub-three year correlation with rainfall series clearly shows the importance of climate in determining inter-annual variability.

Although the cross-wavelet analysis (matching amplitudes of two time series) were not significant, wavelet coherence analysis revealed significant correlation at periodicities more than one year (chapters 3 and chapter 5). The wavelet coherence analysis (phase synchrony concept) is often advocated (Cazelles, et al., 2007, Hurtado et al 2014) to establish correlation between non-stationary time series'. The two signals are said to be phase synchronized if their respective phases lock together. The amplitude (cross-wavelet) of the two signals may not necessarily be synchronized or correlated with each other. Using the phase synchrony concept (wavelet coherence analysis) (Rosenblum et al., 1996; Pikovsky et al., 1997, 2001) it is possible to detect weak correlations between non-stationary time series. Therefore, the wavelet coherence results presented in this analysis are not surprising considering the non-stationary pattern in the rainfall time series (Fig. 5.6 in Chapter5).

The majority of temporal analyses of vector-borne diseases focus on developing forecasting models and quantifying the role of environmental variables. There are very few studies which have analysed time series data to tease apart the role of intrinsic and extrinsic factors in driving the outbreaks or cases (Koelle & Pascual, 2004; Koelle et al., 2005; Stenseth et al., 2006). The use of popular ARIMA models is more focussed on forecasting and not used to understand intrinsic mechanisms (Promprou et al., 2006). GLM models are also employed in many studies, but ignore the temporal correlation effect in the dependent variable. The use of GLMM is gaining more importance to account for temporal autocorrelation, but these models have not been used to disentangle the role of intrinsic and extrinsic factors in a Bayesian framework. Recently wavelet methods (Cazelles et al., 2005; Grenfell et al., 2001) have been used to understand the periodicity in the dependent variable and also to identify correlation (wavelet coherence spectra) between two non-stationary time series. Although Bayesian GLMM models offer flexibility in specifying priors and in borrowing information from the points which are nearest in time, and the incorporation of AR(1) can also account for missing covariates (covariates which have not been included in the analysis) with temporal structure, such models are less often applied in analysing infectious disease data. Bayesian GLMM models applied to outbreak data (Chapter 3) can account for uncertainty in the parameter values and can be very useful in making forecasting models.

7.3.2 Role of Intrinsic and extrinsic factors in spatial variability of BTV outbreaks

Spatial variation in BTV outbreaks can be due to either intrinsic and extrinsic factors or both. Spatial analysis using a Poisson model with spatial autocorrelation (BYM approach) and covariates outperformed the models with only covariates (chapter 4). Spatial autocorrelation as discussed in chapter 4 can be due to spread of disease through movement of infected vectors or host animals or due to missing covariates with spatial structure and both can be grouped in the extrinsic factor category. The spatial variability in BTV outbreaks at district level can be due to innate resistance or susceptibility (intrinsic factors) to BTV outbreaks as discussed in chapter 1. Outperformance of models with host variables (breed type and abundance) demonstrates the role of intrinsic factors in comparison to land cover and climate (extrinsic factors). The individual category of models without spatial autocorrelation performed worse than the models with spatial autocorrelation. Overall, the intrinsic factors dominate over the extrinsic factors in determining the spatial variability of BTV outbreaks.

Spatial analysis is very common in non-infectious disease mapping and it is gaining importance in infectious disease epidemiology. The potential of Bayesian GLMM models to account for spatial autocorrelation and the identification of risk factors via the all subset approach in INLA has not yet been fully realised. A pentavalent vaccination campaign has recently been launched, currently targeted only at sheep. The risk factors identified in the present spatial analyses include other hosts (buffalo and cattle) and these should also be covered by the vaccination programme and in sero-surveillance. In many BTV affected countries, the vaccination is practiced in both cattle and sheep (Caporale, & Giovannini, 2010), but there are no reports of vaccination in buffalo.

7.3.3 Space-time methods to understand intrinsic and extrinsic factors

District level annual outbreaks of BTV analysed using Bayesian Poisson regression model by accounting for extra-Poisson variability was helpful in teasing apart the relative role of intrinsic and extrinsic factors. The direct influence of sea surface temperature on the bluetongue outbreaks is difficult to establish using wavelet analysis tools due to the relatively short time series of BTV data, however the longer time series of both rainfall (52 years) and sea surface temperature was helpful in establishing the strong correlation between North-East monsoon rainfall and sea surface temperature for all the three states. The importance of rainfall in determining the severity of BTV outbreaks was established in both the state level time series model (Chapter 4) and of both rainfall and temperature in district level space-time analysis (Chapter 5).

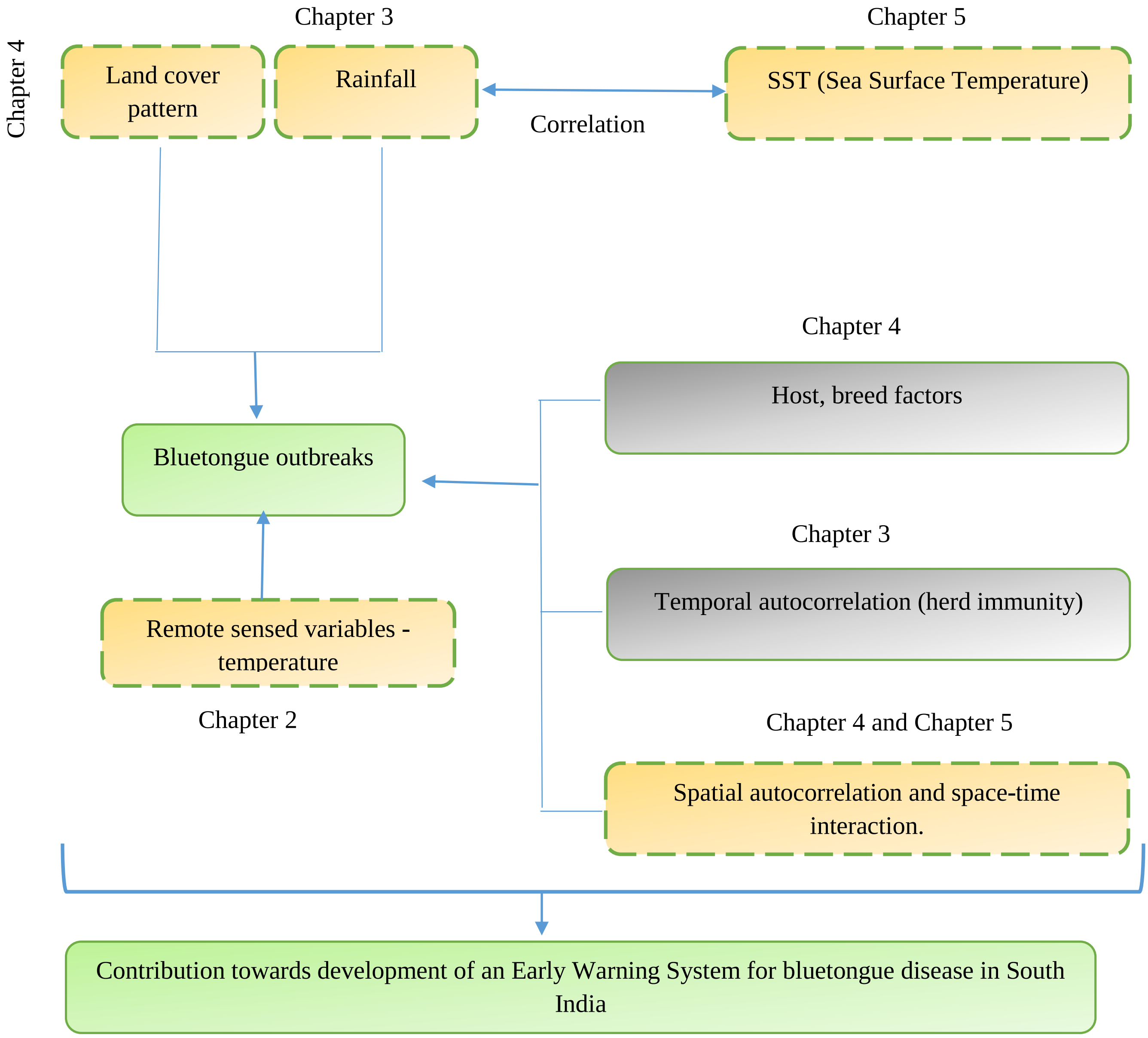

Chapter 3
Chapter 5
Chapter 4
Land cover pattern
Rainfall
SST (Sea Surface Temperature)
Correlation
Chapter 4
Host, breed factors
Bluetongue outbreaks
Chapter 3
Temporal autocorrelation (herd immunity)
Remote sensed variables - temperature
Chapter 2
Chapter 4 and Chapter 5
Spatial autocorrelation and space-time interaction.
Contribution towards development of an Early Warning System for bluetongue disease in South India

Figure 7.1: Influence of different intrinsic and extrinsic factors on bluetongue outbreaks in South India across different spatial and temporal scales. Boxes with dotted lines indicate extrinsic factors and boxes with solid line indicate intrinsic factors.

7.4 Variable selection

Variable selection in analyses of epidemiological data is critical. In this study, different variable selection methods were employed in different chapters depending on the objectives of the analyses and also the computational feasibility to understand the epidemiology of BTV and develop predictive models. Step-wise variable selection was employed in NLDA (Chapter 2) as the aim of the analysis was to identify variables which discriminate between presence and absence groups.

In the time series analysis, MCMC based simulation was performed and a long time (e.g. around 45 minutes) was required to run multiple chains with large numbers of iterations (200,000) for a single model. Therefore, stage wise selection of models (entering monthly variables in different stages) using temperature and rainfall variables was performed and the consistency of selected variables was compared when the order of the stages was varied. Nevertheless, the objective of this analysis was to quantify the role of intrinsic and extrinsic factors in determining the seasonality and inter-annual variability of BTV in Andhra Pradesh and identifying significant lags of rainfall and temperature, so a limited number of models were considered.

In the district level spatial analyses of BTV outbreaks, a modified all subsets approach was followed (chapter 4). The modified all subsets approach was possible due to computational efficiency of INLA and there was no bias in selection of variables due to accounting for spatial autocorrelation. Variable selection that ignores spatial autocorrelation can result in the selection of unimportant variables (Hoeting et al., 2006) and also bias in the estimates of the co-efficients as discussed in chapter 4. The all subsets approach in a Bayesian framework has not so far been applied to bluetongue or any other vector borne disease.

Village level analysis of bluetongue cases to identify risk factors and develop predictive model has not previously been conducted for bluetongue or any other animal vector borne disease in India. Bayesian Network Modelling (BNM) was employed in a village level analysis of bluetongue cases, using a total of 21 variables, to identify their direct and indirect relationships (chapter 6). Only those variables that were directly linked to the BTV cases were considered in the subsequent village level models.

7.5 Predictions at different spatial and temporal scales to help in controlling BTV in South India

7.5.1 Global early warning systems for livestock diseases

Early warning systems for animal diseases are often based on collection of data (formal and informal), analysis and the creation of alerts usually based on cumulative case numbers, often seasonally adjusted. Although the alerts are generated in real time, there is no prediction of outbreaks in unknown areas, or quantification of the risk factors involved in disease transmission. The alerts are disseminated through fax, e-mail, bulletins to the member countries. GLEWS (Global Early Warning System) (FAO), is a joint initiative of OIE, WHO and FAO for early warning of important livestock diseases and zoonotic diseases. EMPRES (Emergency preparation system) for Tran’s boundary Animal and Plant Pests and Diseases (Welte & Terán, 2004) was developed by FAO and mainly focuses on Rinderpest, Contagious Bovine Pleuropneumonia (CBPP), FMD (Foot and Mouth Disease), PPR (Peste des Petits Ruminants), Rift valley fever, Newcastle disease, lumpy skin disease and African swine fever. Bluetongue is not included in the list. The EMPRES early warning system is based on reports and the information is disseminated to member countries. There is an early warning system for RVF in Africa (Anyamba et al., 2009) and the system is used by EMPRES for enhanced surveillance activities (of humans and animals). However, the early warning system relies on positive anomalies of NDVI and rainfall to develop risk maps and does not involve any predictive modelling methods. Thus, neither GLEWS nor EMPRES focuses on predictive models or the identification of risk factors.

7.5.2 NADRES model for forecasting bluetongue in India

Apart from understanding the role of intrinsic and extrinsic factors in determining the severity of BTV outbreaks at different spatial and temporal scales, statistical models also help in making predictions in unknown areas and times. These predictions help to inform policy makers for timely control of disease by vector abatement or vaccination. Currently, there is a system of forecasting bluetongue and other livestock diseases (a total of fifteen diseases) in India. The forecasting model developed by (Sudhindra & Rajasekhar 1997) uses logistic regression and all diseases (fifteen economically important livestock diseases) are forecast two months in advance and not updated by retraining the model with new data or new methods. Presence and absence forecasting is done at district level over the whole of India. Control of bluetongue using predictions (presence and absence) at district level is very difficult considering the size of each district and the large number of villages involved (average = 900 villages per district in three states of South India) and the diversity of landscape, climate and host conditions. Bluetongue occurs with varying severity in each district and there is inter-annual variability in South India. Thus presence and absence predictions will be of very little help in planning control measures (vaccination or vector control). Therefore, it is important to have predictive models (predicting outbreak numbers) at various spatial and temporal scales for effective management of bluetongue in South India.

7.5.3 Temporal predictions

Temporal analysis of bluetongue outbreaks quantified the role of climate in determining temporal variability of bluetongue in Andhra Pradesh and a monthly forecasting model developed, which can help disease managers to employ veterinary personnel and other resources to control the disease in a timely manner (Chapter 3). The forecasting model developed at the state level will be very helpful in predicting hyper endemic years as the model captures both endemic and hyper endemic years, but needs further improvement to be used in early warning system by incorporating other seasonal variables such as relative humidity and wind speed and fine scale rainfall and temperature data. The model was developed using past outbreak data (no vaccination was carried out in the past). Therefore, the current time series model has to be updated in the next few years when vaccination is carried out and also using finer resolution climate data. The time series modelling framework can also be applied to other two endemic states with different seasonality and inter-annual variability in outbreaks. Temporal predictive models in the past were based on linear regression ignoring temporal autocorrelation. ARIMA or SARIMA are not suitable for count data as discussed in the introduction (Chapter 1). Bayesian time series models accounting for temporal autocorrelation have not been applied to bluetongue data to date.

7.5.4 District level spatial analysis of bluetongue outbreaks

District level spatial risk maps (Chapter 4) will be very helpful to the disease managers to plan and utilise their limited resources to control the disease in high risk areas using vaccination and vector control measures. The methodology developed can also be extended to analyse other livestock diseases which are causing huge economic losses to the country and also affecting the livelihood of farmers. Overall the spatial risk of bluetongue at district level is attributed to host and land cover predictor variables and the inter-annual variability in bluetongue outbreaks is determined by climatic variation (both temperature and rainfall).

7.5.5 District level annual predictions of BTV outbreaks

Currently, there is no system of forecasting outbreak numbers (as opposed to presence) for any livestock disease in India as the NADRES system discussed earlier predicts presence and absence of a disease at district level. District level annual predictions (Chapter 5) using Bayesian model by accounting for spatial and temporal autocorrelation will be effective to plan the control measures well in advance as the bluetongue season starts in the month of September giving effective lead time for the disease managers.

The correlation between observed and predicted BTV outbreaks for the year 2009 was poor, but the training model shows very high correlation (r^2= 0.99) and the cross validation statistics (CPO) are encouraging for the modelling approach to be used, alongside improved environmental covariates, in forecasting BTV outbreaks at district level.

7.5.6 Lead times for forecasts

The lead times for the temporal and spatio-temporal predictive models depend on the requirements of the end users. The forecasting system should constantly be updated by improving surveillance for early detection and also incorporating additional data on both climatic and non-climatic factors. Biologically, the feasible lead time that can be built into early warning frameworks also depends on the life history parameters of the potential vectors for bluetongue. The time series forecast (state level) using rainfall at lag 2 can be effectively used to forecast the disease two months in advance for the state level models (Chapter 3). In the district level annual forecast (Chapter 5) developed, the lead time (approximately 6 months until it is disseminated to the stakeholders), which can be effective for the disease managers to channelize their resources for vector control or vaccination.

Bayesian methods offers advantages over traditional frequentist methods by accounting for unobserved or unavailable predictor variables and thus the regression co-efficients obtained are reliable and can be used for developing predictive models. Use of Bayesian methods is gaining importance in public health, but the focus is more on non-infectious diseases. Use of such methods is not very common in veterinary Epidemiology.

7.5.7 Stakeholders using risk maps and predictive models

The predictive models developed in this thesis will be helpful to disease managers at different hierarchical levels. Generally, the current system of veterinary and animal husbandry department in South India operates at broadly three levels. The Animal Husbandry director (state level) is the chief officer and head of the department. At the second level is the district level officer, who is responsible for the district level activities. At the third level is the Veterinary Officer who works at the village level. The state level officer is mainly responsible for taking any decision with respect to disease control measures and is supported by Joint directors for each department (extension, animal health, statistics). The district level officer looks after the district level activities and monitors the veterinary officers for successful implementation of any policy (vaccination, extension activities or implementing any scheme). The veterinary officers working in different villages are responsible for treating of animals, vaccination, de-worming, and extension activities. The veterinary officer is the person who actually reports the disease which is later on compiled at district level.

The state level time series models (Chapter 3) and district level risk maps (Chapter 4) will be helpful to state veterinary officers to plan and allocate resources (vaccinations, personnel). The district level officers will be informed by the state level officer about the yearly district level forecast (Chapter 5) to plan ahead for vaccinations, health camps, and extension activities to promote vaccination by the villagers in case of higher risk predictions for that particular year and routine vaccination if there is low to moderate risk of bluetongue for that particular year. The village level predictions (both presence & absence and case numbers) will be helpful to both district level officers and veterinary officers to plan vaccination in higher risk areas on priority basis and also creating awareness among the farmers. The targeted vaccination at village level will not only help in controlling the disease, but also order number of doses required to avoid wastage of vaccines. The vaccines are normally provided based on the livestock population and there is no risk map or predictions at village level for ordering vaccine doses.

7.5.8 Risk communication to stakeholders

Risk maps and predictive models developed at different spatial and temporal scales need to be disseminated to stakeholders. The static risk maps (presence & absence or case numbers) can be disseminated in the form of leaflets or maps.   The presence and absence risk maps (chapter 2) can be provided to different stakeholders for different purposes. The state level authorities can be informed about the bluetongue risks in their respective state. The village level predictions of case numbers can be provided to the district level officers in the table format (list of villages with predictions of case numbers), who in turn can inform the village level Veterinary officers to carry out vaccinations and awareness campaigns in high risk areas. The state level time series predictions can be disseminated in the form of bulletins two months in advance to the state level officers for planning the control measures, along with annual district level predictions. The risk maps and predictions can also be provided in the website form for the stakeholders along with the instructions on how to use such maps and predictions. The risk maps and predictions at different spatial and temporal scales can also be used by the researchers and epidemiologist working on bluetongue to plan surveillance and vector surveys. The interaction between the stakeholders, epidemiologist and researchers at regular intervals can help improve the risk maps and predictions.

7.6 Summary and conclusions

The analyses presented in different chapters help to better understand the epidemiology of bluetongue in South India. Although different suites of risk factors (satellite derived variables, host, land cover and climate) were considered in the analyses, additional data on herd immunity, farm level factors, information on direction and distance on migration of sheep and other socio-economic factors will help in better explaining the structured and unstructured heterogeneity in the bluetongue outbreak data. The predictive models developed in this thesis at various spatial and temporal scales can be helpful in controlling BTV. Although the space-time "out of fit" forecast does not show close correspondence with the observed outbreaks, but the training model shows good (99%) correspondence with observed outbreaks and, in combination with predictive models at other spatial (district level and village level) and temporal (state level) scales, may eventually contribute to an Early Warning System for bluetongue in India, but there is clearly much more work to be done. The model framework presented here can be supplemented with data on vaccination, immunity, vector abundance and socio-economic factors. When the time comes, accurate predictions can be disseminated to the stakeholders by means of website, leaflets, and bulletins.

References

Acevedo, P., Ruiz-Fons, F., Estrada, R., Márquez, A. L., Miranda, M. A., Gortázar, C. & Lucientes, J. (2010). A broad assessment of factors determining *Culicoides imicola* abundance: modelling the present and forecasting its future in climate change scenarios. *PLoS One,* ***5***(12), e14236.

Adusumilli, R. & Laxmi, S. B. (2011). Potential of the system of rice intensification for systemic improvement in rice production and water use: the case of Andhra Pradesh, India. *Paddy and Water Environment,* ***9***(1), 89-97.

Ahuja, V., George, P. S., Ray, S., Kurup, M. P. G. & Gandhi, V. P. (2000). Agricultural services and the poor: Case of livestock health and breeding services in India.

Ahuja, V., Rajasekhar, M. & Raju, R. (2008). Animal health for poverty alleviation: A review of key issues for India. *Background paper prepared for livestock sector review of the World Bank*.

Alexander, N., Moyeed, R. & Stander, J. (2000). Spatial modelling of individual-level parasite counts using the negative binomial distribution. *Biostatistics,* ***1***(4), 453-463.

Ali, J. (2007). Livestock sector development and implications for rural poverty alleviation in India. *Livestock Research for Rural Development,* ***19***(2), 1-15.

Altizer, S., Dobson, A., Hosseini, P., Hudson, P., Pascual, M. & Rohani, P. (2006). Seasonality and the dynamics of infectious diseases. *Ecology Letters,* ***9***(4), 467-484.

Anderson, R. M., May, R. M. & Anderson, B. (1992). *Infectious diseases of humans: dynamics and control* (Vol. 28): Wiley Online Library.

Annamalai, H., Hamilton, K. & Sperber, K. R. (2007). The South Asian summer monsoon and its relationship with ENSO in the IPCC AR4 simulations. *Journal of Climate*, ***20***(6), 1071-1092.

Anyamba, A., Chretien, J.P., Small, J., Tucker, C.J., Formenty, P.B., Richardson, J.H., Britch, S.C., Schnabel, D.C., Erickson, R.L. & Linthicum, K.J. (2009). Prediction of a Rift Valley fever outbreak.*Proceedings of the National Academy of Sciences*, ***106***(3), pp.955-959.

Arun, S., John, K., Ravishankar, C., Mini, M., Ravindran, R. & Prejit, N. (2014). Seroprevalence of Bluetongue among domestic ruminants in Northern Kerala, India. *Tropical Biomedicine,* ***31***(1), 26-30.

Assunção, R. M., Reis, I. A. & Oliveira, C. D. L. (2001). Diffusion and prediction of Leishmaniasis in a large metropolitan area in Brazil with a Bayesian space–time model. *Statistics in Medicine*, ***20***(15), 2319-2335.

Babyak, M. A. (2004). What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine,* ***66***(3), 411-421.

Bandhyopadhyay, S. K. & Mallick, B. B. (1983). Serological prevalence of bluetongue antibodies in India. *Indian Journal of Animal Sciences*.

Banerjee, S. & Fuentes, M. (2012). Bayesian modeling for large spatial datasets. *Wiley Interdisciplinary Reviews: Computational Statistics,* ***4***(1), 59-66.

Banerjee, S., Gelfand, A. E. & Carlin, B. P. (2004). *Hierarchical modeling and analysis for spatial data*: Crc Press.

Banerjee, S., Gelfand, A. E., Finley, A. O. & Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), **70***(4), 825-848.

Batten, C. A., Maan, S., Shaw, A. E., Maan, N. S. & Mertens, P. P. (2008). A European field strain of bluetongue virus derived from two parental vaccine strains by genome segment reassortment. *Virus Research*, **137**(1), 56-63.

Baylis, M., Bouayoune, H., Touti, J. & El Hasnaoui, H. (1998). Use of climatic data and satellite imagery to model the abundance of Culicoides imicola, the vector of African horse sickness virus, in Morocco. *Medical and Veterinary Entomology, **12***(3), 255-266.

Baylis, M., Meiswinkel, R. & Venter, G. (1999). A preliminary attempt to use climate data and satellite imagery to model the abundance and distribution of Culicoides imicola (Diptera: Ceratopogonidae) in southern Africa. *Journal of the South African Veterinary Association, **70***(2), p. 80-89.

Baylis, M., Mellor, P. S. & Meiswinkel, R. (1999b). Horse sickness and ENSO in South Africa. *Nature, **397***(6720), 574-574.

Baylis, M., Mellor, P., Wittmann, E. & Rogers, D. (2001). Prediction of areas around the Mediterranean at risk of bluetongue by modelling the distribution of its vector using satellite imaging. *The Veterinary Record, **149***(21), 639-643.

Baylis, M., O’Connell, L. & Purse, B. V. (2004). Modelling the distribution of bluetongue vectors. *Veterinaria Italiana, **40***(3), 176-181.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M. & Songini, M. (1995). Bayesian analysis of space—time variation in disease risk. *Statistics in Medicine*, ***14***(21-22), 2433-2443.

Besag, J., & Newell, J. (1991). The detection of clusters in rare diseases. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 143-155.

Besag, J., York, J. & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics, **43***(1), 1-20.

Best, N., Richardson, S., & Thomson, A. (2005). A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research, **14***(1), 35-59.

Bhalodiya, M. & Jhala, M. (2002). Seroepidemiological study of bluetongue virus using AB-ELISA. *Indian Veterinary Journal (India)*.

Bhanuprakash, V., Saravanan, P., Hosamani, M., Balamurugan, V., Mondal, B. & Singh, R. K. (2007). Status of sheep sera to bluetongue, peste des petits ruminants and sheep pox in a few northern states of India. *Veterinaria Italiana*, ***44***(3), 527-536.

Bhaskaran, K., Gasparrini, A., Hajat, S., Smeeth, L., & Armstrong, B. (2013). Time series regression studies in environmental epidemiology. *International Journal of Epidemiology*, dyt092.

Bhoyar, R., Udupa, K., Reddy, P., Kasaralikar, V. & Prasad, C. (2012). Climatological factors associated with abundance of Culicoides midges. *Journal of Veterinary Parasitology,* ***26***(2), 148-150.

Biggeri, A., Dreassi, E., Catelan, D., Rinaldi, L., Lagazio, C. & Cringoli, G. (2006). Disease mapping in veterinary epidemiology: a Bayesian geostatistical approach. *Statistical Methods in Medical Research,* ***15***(4), 337-352.

Bitew, M., Sukdeb, N., Ravishankar, C. & Somvanshi, R. (2013). Serological and molecular evidence of bluetongue in sheep and goats in Uttar Pradesh, India. *African Journal of Biotechnology,* ***12***(19), 2699-2705.

Bivand, R. S., Pebesma, E. J., Gómez-Rubio, V. & Pebesma, E. J. (2008). *Applied spatial data analysis with R* (Vol. 747248717): Springer.

Blangiardo, M., Cameletti, M., Baio, G. & Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. *Spatial and Spatiotemporal Epidemiology, 7,* 39-55.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H. & White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution,* ***24***(3), 127-135.

Bonneau, K., DeMaula, C., Mullens, B., & MacLachlan, N. (2002). Duration of viraemia infectious to *Culicoides sonorensis* in bluetongue virus-infected cattle and sheep. *Veterinary Microbiology,* ***88***(2), 115-125.

Boyd, D. S. & Curran, P. J. (1998). Using remote sensing to reduce uncertainties in the global carbon budget: the potential of radiation acquired in middle infrared wavelengths. *Remote Sensing Reviews*, ***16***(4), 293-327.

Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88*(421), 9-25.

Brodie, S. J., Wilson, W. C., O'Hearn, P. M., Muthui, D., Diem, K. & Pearson, L. D. (1998). The effects of pharmacological and lentivirus-induced immune suppression on orbivirus pathogenesis: assessment of virus burden in blood monocytes and tissues by reverse transcription-in situ PCR. *Journal of Virology, **72***(7), 5599-5609.

Burgin, L., Gloster, J., Sanders, C., Mellor, P., Gubbins, S. & Carpenter, S. (2013). Investigating Incursions of Bluetongue Virus Using a Model of Long-Distance Culicoides Biting Midge Dispersal. *Transbound Emerging Disease, **60***(3), 263-272.

Burke, D., Carmichael, A., Focks, D. & Grimes, D. (2001). Under the weather: climate, ecosystems, and infectious disease. *Emerging Infectious Diseases, **7***(3).

Burnham, K. P. & Anderson, D. R. (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*(2), 261-304.

Calistri, P., Goffredo, M., Caporale, V. & Meiswinkel, R. (2003). The distribution of *Culicoides imicola* in Italy: application and evaluation of current Mediterranean models based on climate. *Journal of Veterinary Medicine, Series B,* ***50***(3), 132-138.

Campbell, J. B. (2002). *Introduction to remote sensing*. CRC Press.

Caporale, V. & Giovannini, A. (2010). Bluetongue control strategy, including recourse to vaccine: a critical review. *Revue Scientifique et Technique (International Office of Epizootics),* ***29***(3), 573-591.

Carpenter, S., Wilson, A., Barber, J., Veronesi, E., Mellor, P., Venter, G., & Gubbins, S. (2011). Temperature dependence of the extrinsic incubation period of orbiviruses in *Culicoides* biting midges. *PLoS One*, ***6***(11), e27987.

Cazelles, B., Chavez, M., Berteaux, D., Ménard, F., Vik, J. O., Jenouvrier, S. & Stenseth, N. C. (2008). Wavelet analysis of ecological time series. *Oecologia,* ***156***(2), 287-304.

Cazelles, B., Chavez, M., de Magny, G. C., Guégan, J.-F. & Hales, S. (2007). Time-dependent spectral analysis of epidemiological time-series with wavelets. *Journal of The Royal Society Interface,* ***4***(15), 625-636.

Cazelles, B., Chavez, M., McMichael, A. J. & Hales, S. (2005). Nonstationary influence of El Nino on the synchronous dengue epidemics in Thailand. *PLoS Medicine,* ***2***(4), e106.

Chacko, C. T., Gopikrishna, P., Tiwari, S. & Ramesh, V. (2010). Growth, Efficiency Gains, and Social Concerns. *Livestock in a changing landscape: experiences and regional perspectives*, 55.

Chatfield, C. (2013). *The analysis of time series: an introduction*: CRC press.

Chauhan, H., Chandel, B., Vasava, K., Patel, A., Shah, N. & Kher, H. (2004). Seroprevalence of bluetongue in Gujarat. *Indian Journal of Comparative Microbiology, Immunology and Infectious Diseases, **25***(2), 80-83.

Chaves, L. F. & Pascual, M. (2006). Climate cycles and forecasts of cutaneous leishmaniasis, a nonstationary vector-borne disease. *PLoS Medicine, **3***(8), e295.

Chaves, L. F. & Pascual, M. (2007). Comparing models for early warning systems of neglected tropical diseases. *PLoS Neglected Tropical Diseases, **1***(1), e33.

Chou, W.-C., Wu, J.-L., Wang, Y.-C., Huang, H., Sung, F.-C. & Chuang, C.-Y. (2010). Modeling the impact of climate variability on diarrhea-associated diseases in Taiwan (1996–2007). *Science of the Total Environment, **409***(1), 43-51.

Clayton, D. G. (1996). Generalized linear mixed models *Markov chain Monte Carlo in practice* (pp. 275-301): Springer.

Clements, A. C., Moyeed, R. & Brooker, S. (2006). Bayesian geostatistical prediction of the intensity of infection with Schistosoma mansoni in East Africa. *Parasitology, **133***(06), 711-719.

Clements, A., Barnett, A. G., Cheng, Z. W., Snow, R. W. & Zhou, H. N. (2009). Space-time variation of malaria incidence in Yunnan province, China. *Malaria Journal, **8***(180), 10.1186.

Coetzee, P., Stokstad, M., Venter, E. H., Myrmel, M. & Van Vuuren, M. (2012). Bluetongue: a historical and epidemiological perspective with the emphasis on South Africa. *Virology Journal, **9***(198), 1-9.

Congalton, R. G. (1991). Remote sensing and geographic information system data integration: error sources and. *Photogrammetric Engineering & Remote Sensing*, **57**(6), 677-687.

Congdon, P. (2007). *Bayesian statistical modelling* (Vol. 704): John Wiley & Sons.

Conte, A., Giovannini, A., Savini, L., Goffredo, M., Calistri, P. & Meiswinkel, R. (2003). The effect of climate on the presence of *Culicoides imicola* in Italy. *Journal of Veterinary Medicine, Series B, **50***(3), 139-147.

Conte, A., Goffredo, M., Ippoliti, C. & Meiswinkel, R. (2007). Influence of biotic and abiotic factors on the distribution and abundance of *Culicoides imicola* and the Obsoletus Complex in Italy. *Veterinary Parasitology, **150***(4), 333-344.

Dadawala, A. I., Biswas, S. K., Rehman, W., Chand, K., De, A., Mathapati, B. S., Kumar, P., Chauhan, H. C., Chandel, B. S. & Mondal, B. (2012). Isolation of Bluetongue Virus Serotype 1 from *Culicoides* vector Captured in Livestock Farms and Sequence Analysis of the Viral Genome Segment-2. *Transboundary and Emerging Diseases, **59***(4), 361-368. doi:DOI 10.1111/j.1865-1682.2011.01279.x

Dadawala, A. I., Kher, H. S., Chandel, B. S., Bhagat, A. G., Chauhan, H. C., Ranjan, K. & Minakshi, P. (2013). Isolation and Molecular

Characterization of Bluetongue Virus 16 of Goat Origin from India. *Advances in Animal and Veterinary Sciences*, ***1***(4S), 24-29.

Daniels, P., Sendow, I., Pritchard, L. & Eaton, B. (2003). Regional overview of bluetongue viruses in South-East Asia: viruses, vectors and surveillance. *Veterinaria Italiana,* ***40***(3), 94-100.

Defourny, P., Vancutsem, C., Bicheron, P., Brockmann, C., Nino, F., Schouten, L. & Leroy, M. (2006). *GLOBCOVER: a 300 m global land cover product for 2005 using Envisat MERIS time series.* Paper presented at the ISPRS Commission VII Mid-term Symposium "Remote Sensing From Pixels to Processes", Enschede, the Netherlands.

Delgado, C. L., Rosegrant, M. W., Steinfeld, H., Ehui, S. & Courbois, C. (1999). The coming livestock revolution. *Background paper n. 6, Department of Economic and Social Affairs, Commission of Sustainable Development, Eighth Session*.

DeMaula, C. D., Leutenegger, C. M., Bonneau, K. R. & MacLachlan, N. J. (2002a). The role of endothelial cell-derived inflammatory and vasoactive mediators in the pathogenesis of bluetongue. *Virology*, ***296***(2), 330-337.

DeMaula, C. D., Leutenegger, C. M., Jutila, M. A. & MacLachlan, N. J. (2002b). Bluetongue virus-induced activation of primary bovine lung microvascular endothelial cells. *Veterinary Immunology and Immunopathology,* ***86***(3), 147-157.

Desai, P. (2004). Sero-prevalence of bluetongue in cattle in two South Gujarat districts. *Indian Veterinary Journal (India)*.

Diggle, P. J., Ribeiro Jr, P. J. & Christensen, O. F. (2003). An introduction to model-based geostatistics *Spatial Statistics and Computational Methods* (pp. 43-86): Springer.

Diggle, P. J., Tawn, J. & Moyeed, R. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics), **47***(3), 299-350.

Diggle, P., Heagerty, P., Liang, K.-Y. & Zeger, S. (2002). *Analysis of longitudinal data*: Oxford University Press.

Diggle, P., Moyeed, R., Rowlingson, B. & Thomson, M. (2002b). Childhood malaria in the Gambia: a case-study in model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics), **51***(4), 493-506.

Dilley, M. & Heyman, B. N. (1995). ENSO and disaster: droughts, floods and El Nino/Southern Oscillation warm events. *Disasters, **19***(3), 181-193.

Dohoo, I. R., Ducrot, C., Fourichon, C., Donald, A. & Hurnik, D. (1997). An overview of techniques for dealing with large numbers of independent variables in epidemiologic studies. *Preventive Veterinary Medicine, **29***(3), 221-239.

Dormann, C. F., M McPherson, J., B Araújo, M., Bivand, R., Bolliger, J., Carl, G., G Davies, R., Hirzel, A., Jetz, W. & Daniel Kissling, W. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography, **30***(5), 609-628.

Dubay, S. A., deVos Jr, J. C., Noon, T. H. & Boe, S. (2004). Epizootiology of hemorrhagic disease in mule deer in central Arizona. *Journal of Wildlife Diseases, **40***(1), 119-124.

Dungu, B., Gerdes, T. & Smit, T. (2004). The use of vaccination in the control of bluetongue in southern Africa. *Veterinaria Italiana, **40***(4), 616-622.

Eastman, J. R. & Filk, M. (1993). Long sequence time series evaluation using standardized principal components. *Photogrammetric Engineering and Remote sensing, **59***(6), 991-996.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics, **32***(2), 407-499.

Elith, J. & Graham, C. H. (2009). Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. *Ecography, **32***(1), 66-77.

Erasmus, B. (1975). Bluetongue in sheep and goats. *Australian Veterinary Journal, **51***(4), 165-170.

Erasmus, B. J., Potgieter, A., Mertens, P., Bayliss, M. & Mellor, P. (2009). The history of bluetongue. *Bluetongue,* 7-21.

Faes, C., van der Stede, Y., Guis, H., Staubach, C., Ducheyne, E., Hendrickx, G. & Mintiens, K. (2013). Factors affecting Bluetongue serotype 8 spread in Northern Europe in 2006: The geographical epidemiology. *Preventive Veterinary Medicine, **110***(2), 149-158.

FAO, O. WHO. 2011 GLEWS: global early warning and response system for major animal diseases, including zoonoses. *See http://www. glews. net/(accessed 5 January 2011).*

Farnsworth, M. L., Hoeting, J. A., Hobbs, N. T. & Miller, M. W. (2006). Linking chronic wasting disease to mule deer movement scales: a hierarchical Bayesian approach. *Ecological Applications*, ***16***(3), 1026-1036.

Fernández, M. Á. L., Bauernfeind, A., Jiménez, J. D., Gil, C. L., El Omeiri, N. & Guibert, D. H. (2009). Influence of temperature and rainfall on the evolution of cholera epidemics in Lusaka, Zambia, 2003–2006: analysis of a time series. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, ***103***(2), 137-143.

Fielding, A. H. & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental conservation*, **24**(01), 38-49.

Finley, A. O., Banerjee, S. & Carlin, B. P. (2007). spBayes: an R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, ***19***(4), 1.

Finley, A. O., Sang, H., Banerjee, S. & Gelfand, A. E. (2009). Improving the performance of predictive process modeling for large datasets. *Computational Statistics & Data analysis*, ***53***(8), 2873-2884.

Food and Agriculture Organization of the United Nations, FAOSTAT database (FAOSTAT, 2008), available at http://faostat.fao.org.

Friedman, N., Nachman, I. & Peér, D. (1999, July). Learning bayesian network structure from massive datasets: the «sparse candidate «algorithm. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence* (pp. 206-215). Morgan Kaufmann Publishers Inc..

Fuentes, M. (2007). Approximate likelihood for large irregularly spaced spatial data. *Journal of the American Statistical Association, **102***(477), 321-331.

Furnival, G. M. & Wilson, R. W. (1974). Regressions by leaps and bounds. *Technometrics, **16***(4), 499-511.

Gambles, R. (1949). Bluetongue of sheep in Cyprus. *Journal of Comparative Pathology and Therapeutics, **59***, 176-190.

Garcia-Saenz, A., Saez, M., Napp, S., Casal, J., Saez, J. L., Acevedo, P., Guta, S. & Allepuz, A. (2014). Spatio-temporal variability of bovine tuberculosis eradication in Spain (2006–2011). *Spatial and Spatiotemporal Epidemiology, **10***, 1-10.

Gelfand, A. E. (1996). Model determination using sampling-based methods *Markov chain Monte Carlo in practice* (pp. 145-161): Springer.

Gelfand, A. E., Schmidt, A. M., Wu, S., Silander, J. A., Latimer, A. & Rebelo, A. G. (2005). Modelling species diversity through species level hierarchical modelling. *Journal of the Royal Statistical Society: Series C (Applied Statistics), **54***(1), 1-20.

Gelman, A. & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*: Cambridge University Press.

George, E. I. (2000). The variable selection problem. *Journal of the American Statistical Association*, ***95***(452), 1304-1308.

Gharbi, M., Quenel, P., Gustave, J., Cassadou, S., Ruche, G. L., Girdary, L. & Marrama, L. (2011). Time series analysis of dengue incidence in Guadeloupe, French West Indies: forecasting models using climate variables as predictors. *BMC Infectious Diseases,* ***11***(1), 166.

Gibbs, E. P. J. & Greiner, E. C. (1994). The epidemiology of bluetongue. *Comparative Immunology Microbiology Infectious Diseases,* ***17***(3), 207-220.

Greenland, S. (1989). Modeling and variable selection in epidemiologic analysis. *American Journal of Public Health,* ***79***(3), 340-349.

Grenfell, B., Bjørnstad, O. & Kappey, J. (2001). Travelling waves and spatial hierarchies in measles epidemics. *Nature,* ***414***(6865), 716-723.

Grinsted, A., Moore, J. C. & Jevrejeva, S. (2004). Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear Processes in Geophysics,* ***11***(5/6), 561-566.

Gubbins, S., Carpenter, S., Baylis, M., Wood, J. L. & Mellor, P. S. (2008). Assessing the risk of bluetongue to UK livestock: uncertainty and sensitivity analyses of a temperature-dependent model for the basic reproduction number. *Journal of the Royal Society Interface,* ***5***(20), 363-371.

Haining, R. P. (2003). *Spatial data analysis*: Cambridge University Press Cambridge.

Hales, S., Weinstein, P., Souares, Y. & Woodward, A. (1999). El Niño and the dynamics of vectorborne disease transmission. *Environmental Health Perspectives,* ***107***(2), 99.

Harris, I., Jones, P., Osborn, T. & Lister, D. (2014). Updated high-resolution grids of monthly climatic observations–the CRU TS3. 10 Dataset. *International Journal of Climatology,* ***34***(3), 623-642.

Hartemink, N., Purse, B., Meiswinkel, R., Brown, H., De Koeijer, A., Elbers, A., Boender, G.-J., Rogers, D. & Heesterbeek, J. (2009). Mapping the basic reproduction number (R 0) for vector-borne diseases: a case study on bluetongue virus. *Epidemics,* ***1***(3), 153-161.

Hassan, A., Walton, T. & Osburn, B. (1992). *Epidemiology of bluetongue virus infection in Malaysia.* Paper presented at the Bluetongue, African horse sickness, and related orbiviruses: Proceedings of the Second International Symposium.

Hay, S. I., Snow, R. W. & Rogers, D. J. (1998). Predicting malaria seasons in Kenya using multitemporal meteorological satellite sensor data. *Transactions of the Royal Society of Tropical Medicine and Hygiene,* ***92***(1), 12-20.

Heckerman, D. (1998). *A tutorial on learning with Bayesian networks*: Springer.

Heffernan, C., Thomson, K. & Nielsen, L. (2011). Caste, livelihoods and livestock: An exploration of the uptake of livestock vaccination adoption among poor farmers in India. *Journal of International Development,* ***23***(1), 103-118.

Heinen, A. (2003). Modelling time series count data: an autoregressive conditional Poisson model. *Available at SSRN 1117187*.

Helfenstein, U. (1986). Box-jenkins modelling of some viral infectious diseases. *Statistics in Medicine,* ***5***(1), 37-47.

Helfenstein, U. (1991). The use of transfer function models, intervention analysis and related time series methods in epidemiology. *International Journal of Epidemiology,* ***20***(3), 808-815.

Henning, M. W. (1949). Animal diseases in South Africa. *Animal Diseases in South Africa* (2nd Edit).

Hesterberg, T., Choi, N. H., Meier, L. & Fraley, C. (2008). Least angle and ℓ1 penalized regression: A review. *Statistics Surveys,* ***2***, 61-93.

Hii, Y. L., Rocklöv, J., Ng, N., Tang, C. S., Pang, F. Y. & Sauerborn, R. (2009). Climate variability and increase in intensity and magnitude of dengue incidence in Singapore. *Global Health Action, 2*.

Hii, Y. L., Rocklöv, J., Wall, S., Ng, L. C., Tang, C. S. & Ng, N. (2012). Optimal lead time for dengue forecast. *PLoS Neglected Tropical Diseases,* ***6***(10), e1848.

Hii, Y. L., Zhu, H., Ng, N., Ng, L. C. & Rocklöv, J. (2012). Forecast of dengue incidence using temperature and rainfall. *PLoS Neglected Tropical Diseases,* ***6***(11), e1908.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology,* ***25***(15), 1965-1978.

Hoeting, J. A., Davis, R. A., Merton, A. A. & Thompson, S. E. (2006). Model selection for geostatistical models. *Ecological Applications,* ***16***(1), 87-98.

Hurtado, L. A., Cáceres, L., Chaves, L. F. & Calzada, J. E. (2014). When climate change couples social neglect: malaria dynamics in Panamá.*Emerging Microbes & Infections,* ***3***(4), e28.

Hurvich, C. M. & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika,* ***76***(2), 297-307.

Ilango, K. (2006). Bluetongue virus outbreak in Tamil Nadu, southern India: Need to study the Indian biting midge vectors, *Culicoides* Latreille (Diptera: Ceratopogonidae). *Current Science-Bangalore-,* ***90***(2), 163.

International Fund for Agricultural Development (Ed.). (2001). *Rural Poverty Report 2001: The Challenge of Ending Rural Poverty*. Oxford University Press.

Isaaks, E. H. & Srivastava, R. M. (1989). *Applied geostatistics* (Vol. 2): Oxford University Press New York.

Ishii, M., Shouji, A., Sugimoto, S. & Matsumoto, T. (2005). Objective analyses of sea-surface temperature and marine meteorological variables for the 20th century using ICOADS and the Kobe collection. *International Journal of Climatology,* ***25***(7), 865-879.

Jain, N., Gupta, Y., Prasad, G., Walton, T. & Osburn, B. (1992). *Bluetongue virus antibodies in buffaloes and cattle in Haryana state of India.* Paper presented at the Bluetongue, African horse sickness, and related orbiviruses: Proceedings of the Second International Symposium.

Kaboli, M., Guillaumet, A. & Prodon, R. (2006). Avifaunal gradients in two arid zones of central Iran in relation to vegetation, climate, and topography. *Journal of Biogeography,* ***33***(1), 133-144.

Kakker, N., Prasad, G., Srivastava, R. & Bhatnagar, P. (2002). Sero-prevalence of bluetongue virus infection in cattle in Haryana, Himachal Pradesh, Punjab and Rajasthan. *Indian Journal of Comparative Microbiology Immunology and Infectious diseases.,* ***23***(2), 147-151.

Kalluri, S., Gilruth, P., Rogers, D. & Szczur, M. (2007). Surveillance of arthropod vector-borne infectious diseases using remote sensing techniques: a review. *PLoS Pathogens*, ***3***(10), 1361-1371.

Keitt, T. H., Bjørnstad, O. N., Dixon, P. M. & Citron-Pousty, S. (2002). Accounting for spatial pattern when modeling organism-environment interactions. *Ecography,* ***25***(5), 616-625.

Kitron, U. (1998). Landscape ecology and epidemiology of vector-borne diseases: tools for spatial analysis. *Journal of Medical Entomology*, ***35***(4), 435-445.

Kitron, U. (2000). Risk maps: transmission and burden of vector-borne diseases. *Parasitology Today*, ***16***(8), 324-325.

Knorr-Held, L. & Besag, J. (1998). Modelling risk from a disease in time and space. *Statistics in Medicine*, ***17***(18), 2045-2060.

Knorr-Held, L. (1999). Bayesian modelling of inseparable space-time variation in disease risk.

Koelle, K. & Pascual, M. (2004). Disentangling extrinsic from intrinsic factors in disease dynamics: a nonlinear time series approach with an application to cholera. *The American Naturalist,* ***163***(6), 901-913.

Koelle, K., Rodó, X., Pascual, M., Yunus, M. & Mostafa, G. (2005). Refractory periods and climate forcing in cholera dynamics. *Nature,* ***436***(7051), 696-700.

Koivisto, M. & Sood, K. (2004). Exact Bayesian structure discovery in Bayesian networks. *The Journal of Machine Learning Research,* **5**, 549-573.

Kovats, R. S. (2000). El Niño and human health. *Bulletin of the World Health Organization,* ***78***(9), 1127-1135.

Kriegel, H. P., Kröger, P., Sander, J. & Zimek, A. (2011). Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* ***1***(3), 231-240.

Kulldorff, M., Mostashari, F., Duczmal, L., Katherine Yih, W., Kleinman, K. & Platt, R. (2007). Multivariate scan statistics for disease surveillance. *Statistics in Medicine*, ***26***(8), 1824-1833.

Kumar, K. R., Pant, G., Parthasarathy, B. & Sontakke, N. (1992). Spatial and subseasonal patterns of the long-term trends of Indian summer monsoon rainfall. *International Journal of Climatology,* ***12***(3), 257-268.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

Lasinio, G. J., Mastrantonio, G. & Pollice, A. (2013). Discussing the "big n problem". *Statistical Methods & Applications,* ***22***(1), 97-112.

Lawson, A. B. (2013). *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*: CRC Press.

Lewis, F. I. & McCormick, B. J. (2012). Revealing the complexity of health determinants in resource-poor settings. *American Journal of Epidemiology*, kws183.

Lewis, F. I. & Ward, M. P. (2013). Improving epidemiologic data analyses through multivariate regression modelling. *Emerging Themes in Epidemiology,* ***10***(1), 4.

Lindgren, F., Rue, H. & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* ***73***(4), 423-498.

Lonkar, P., Uppal, P., Belwal, L. & Mathur, P. (1983). Bluetongue in sheep in India. *Tropical Animal Health and Production,* ***15***(2), 86-86.

Lu, L., Lin, H., Tian, L., Yang, W., Sun, J. & Liu, Q. (2009). Time series analysis of dengue fever and weather in Guangzhou, China. *BMC Public Health,* ***9***(1), 395.

Luz, P. M., Mendes, B. V., Codeço, C. T., Struchiner, C. J. & Galvani, A. P. (2008). Time series analysis of dengue incidence in Rio de

Janeiro, Brazil. *The American Journal of Tropical Medicine and Hygiene*, ***79***(6), 933-939.

Maan, N. S., Maan, S., Belaganahalli, M. N., Ostlund, E. N., Johnson, D. J., Nomikou, K. & Mertens, P. P. (2012). Identification and differentiation of the twenty six bluetongue virus serotypes by RT–PCR amplification of the serotype-specific genome segment 2. *PLoS One*, ***7***(2), e32601.

Maan, S., Maan, N. S., Nomikou, K., Anthony, S. J., Ross-Smith, N., Singh, K. P. & Mertens, P. P. C. (2009). Molecular epidemiology studies of bluetongue virus. *Bluetongue Ist ed*. London: Elsevier, 135-156.

Maan, S., Maan, N. S., Singh, K. P., Belaganahalli, M. N., Guimera, M., Pullinger, G., Nomikou, K. & Mertens, P. P. (2012). Complete genome sequence analysis of a reference strain of bluetongue virus serotype 16. *Journal of Virology*, ***86***(18), 10255-10256.

Maan, S., Maan, N., Samuel, A., Rao, S., Attoui, H. & Mertens, P. (2007). Analysis and phylogenetic comparisons of full-length VP2 genes of the 24 bluetongue virus serotypes. *Journal of General Virology*, ***88***(2), 621-630.

Maclachlan, N., Drew, C., Darpel, K. & Worwa, G. (2009). The pathology and pathogenesis of bluetongue. *Journal of Comparative Pathology*, ***141***(1), 1-16.

MacNab, Y. C., Farrell, P. J., Gustafson, P. & Wen, S. (2004). Estimation in Bayesian disease mapping. *Biometrics*, ***60***(4), 865-873.

Martínez-Beneito, M., López-Quilez, A. & Botella-Rocamora, P. (2008). An autoregressive approach to spatio-temporal disease mapping. *Statistics in Medicine, **27***(15), 2874-2889.

McCarthy, M. A. (2007). *Bayesian Methods for Ecology*: Cambridge University Press.

McCormick, B., Sanchez-Vazquez, M. & Lewis, F. (2013). Using Bayesian networks to explore the role of weather as a potential determinant of disease in pigs. *Preventive Veterinary Medicine*, ***110***(1), 54-63.

McLeod, R., & Kristjanson, P. (1999). Tick cost: economic impact of ticks and TBD to livestock in Africa, Asia and Australia. International Livestock Research Institute (ILRI), Nairobi, Kenya.

McPherson, J. A. N. A., Jetz, W. & Rogers, D. J. (2004). The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact?. *Journal of Applied Ecology*, ***41***(5), 811-823.

Medina, D. C., Findley, S. E., & Doumbia, S. (2008). State–space forecasting of Schistosoma haematobium time-series in Niono, Mali. *PLoS Neglected Tropical Diseases, **2***(8), e276.

Meiswinkel, R. & Braack, L. E. (1994). African horsesickness epidemiology: five species of *Culicoides* (Diptera: Ceratopogonidae) collected live behind the ears and at the dung of the African elephant in the Kruger National Park, South Africa. *The OnderstepoortJjournal of Veterinary Research*, ***61***(2), 155-170.

Mellor, P. & Prrzous, G. (1979). Observations on breeding sites and light-trap collections of *Culicoides* during an outbreak of bluetongue in Cyprus. *Bulletin of Entomological Research, **69***(02), 229-234.

Mellor, P. (2000). Replication of arboviruses in insect vectors. *Journal of Comparative Pathology, **123***(4), 231-247.

Mellor, P., Boned, J., Hamblin, C., & Graham, S. (1990). Isolations of African horse sickness virus from vector insects made during the 1988 epizootic in Spain. *Epidemiology & Infection*, ***105***(02), 447-454.

Mellor, P., Boorman, J. & Baylis, M. (2000). *Culicoides* biting midges: their role as arbovirus vectors. *Annual Review of Entomology, **45***(1), 307-340.

Miller, A. (2002). *Subset selection in regression*: CRC Press.

Mondal, B., Sen, A., Chand, K., Biswas, S., De, A., Rajak, K. & Chakravarti, S. (2009). Evidence of mixed infection of peste des petits ruminants virus and bluetongue virus in a flock of goats as confirmed by detection of antigen, antibody and nucleic acid of both the viruses. *Tropical Animal Health and Production*, ***41***(8), 1661-1667.

Muller, M., Standfast, H., St George, T. & Cybinski, D. (1982). *Culicoides brevitarsis (Diptera: Ceratopogonidae) as a vector of arboviruses in Australia.* Paper presented at the Arbovirus research in Australia Proceedings 3rd symposium.

Myers, M., Rogers, D., Cox, J., Flahault, A. & Hay, S. (2000). Forecasting disease risk for increased epidemic preparedness in public health. *Advances in Parasitology, **47***, 309-330.

Narladkar, B., Deshpande, P. & Shivpuje, P. (2006). Bionomics and life cycle studies on *Culicoides* sp.(Diptera: Ceratopogonidae)*. *Journal of Veterinary Parasitology, **20***(1), 7-12.

Nelder, J. A. & Baker, R. (1972). Generalized linear models. *Encyclopedia of Statistical Sciences*.

Nevill, E. M. (1967). Biological studies on some South African *Culicoides* species (Diptera: Ceratopogonidae) and the morphology of their immature stages.

Nevill, H., Venter, G. J., Meiswinkel, R. & Nevill, E. M. (2007). Comparative descriptions of the pupae of five species of the *Culicoides imicola* complex (Diptera, Ceratopogonidae) from South Africa. *Onderstepoort Journal of Veterinary Research, **74***(2), 97-114.

Onozuka, D. (2014). Effect of non-stationary climate on infectious gastroenteritis transmission in Japan. *Scientific Reports, 4*.

Palaniyandi, M. (2012). The role of remote sensing and GIS for spatial prediction of vector-borne diseases transmission: a systematic review.

Paré, J., Carpenter, T. & Thurmond, M. (1996). Analysis of spatial and temporal clustering of horses with Salmonella krefeld in an intensive care unit of a veterinary hospital. *Journal of the American Veterinary Medical Association, **209***(3), 626-628.

Patel, A., Chandel, B., Chauhan, H., Pawar, D., Bulbule, N., Bhalodia, S. & Kher, H. (2007). Prevalence of potential vector of Bluetongue virus in Gujarat. *Royal Veterinary Journal of India, **3***(1), 33-36.

Patnayak, B.C. (1988). Sheep production and development in India; In sheep production in Asia, In: Devendra, C. and Faylon, P.S (eds.) Proceedings of the workshop on sheep production in Asia, Philippines

Pebesma, E. (2012). spacetime: Spatio-temporal data in r. *Journal of Statistical Software, **51***(7), 1-30.

Pikovsky, A. S. & Kurths, J. (1997). Coherence resonance in a noise-driven excitable system. *Physical Review Letters, **78***(5), 775.

Pikovsky, A., Popovych, O. & Maistrenko, Y. (2001). Resolving clusters in chaotic ensembles of globally coupled identical oscillators. *Physical Review Letters, **87***(4), 044102.

Prasad, G., Jain, N. & Gupta, Y. (1992). Bluetongue virus infection in India: a review. *Revue Scientifique et Technique (International Office of Epizootics), **11***(3), 699-711.

Prasad, G., Sreenivasulu, D. & Singh, K. (2009). Bluetongue in the Indian subcontinent, p 167–196. InMellor PS, Baylis M, Mertens PPC (ed), Bluetongue: Elsevier/Academic Press, London, United Kingdom.

Promprou, S., Jaroensutasinee, M. & Jaroensutasinee, K. (2006). Forecasting dengue haemorrhagic fever cases in Southern Thailand using ARIMA Models. *Dengue Bulletin,* ***30***, 99.

Purse, B. V., Mellor, P. S., Rogers, D. J., Samuel, A. R., Mertens, P. P. & Baylis, M. (2005). Climate change and the recent emergence of bluetongue in Europe. *Nature Reviews Microbiology,* ***3***(2), 171-181.

Purse, B., Baylis, M., Tatem, A., Rogers, D., Mellor, P., Van Ham, M., Chizov-Ginzburg, A. & Braverman, Y. (2004). Predicting the risk of bluetongue through time: climate models of temporal patterns of outbreaks in Israel. *Revue Scientifique et Technique (International Office of Epizootics),* ***23***(3), 761-775.

Purse, B., Carpenter, S., Venter, G., Bellis, G. & Mullens, B. (2015). Bionomics of Temperate and Tropical *Culicoides* Midges: Knowledge Gaps and Consequences for Transmission of *Culicoides*-Borne Viruses. *Annual Review of Entomology,* ***60***, 373-392.

Purse, B., Falconer, D., Sullivan, M., Carpenter, S., Mellor, P., Piertney, S., Albon, S., Gunn, G. & Blackwell, A. (2012). Impacts of climate, host and landscape factors on *Culicoides* species in Scotland. *Medical and Veterinary Entomology,* ***26***(2), 168-177.

Purse, B., Tatem, A., Caracappa, S., Rogers, D., Mellor, P., Baylis, M. & Torina, A. (2004). Modelling the distributions of *Culicoides* bluetongue virus vectors in Sicily in relation to satellite-derived climate variables. *Medical and Veterinary Entomology,* ***18***(2), 90-101.

Purse, B.V., Caracappa, S., Marino, A.M.F., Tatem, A.J., Rogers, D.J., Mellor, P.S., Baylis, M. & Torina, A. (2004). Modelling the distribution of outbreaks and *Culicoides* vectors in Sicily: towards predictive risk maps for Italy. *Veterinaria Italiana,* ***40***(3), pp.303-310.

R Development Core Team (2012) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3- 900051-07-0.

Rao, K. A., Rao, K. S., Rao, S. J., Ravi, A. & Anitha, A. (2013). Analysis of Sheep production systems: North Coastal zone of Andhra Pradesh.

Rao, P., Hegde, N., Reddy, Y., Krishnajyothi, Y., Reddy, Y., Susmitha, B., Gollapalli, S., Putty, K. & Reddy, G. (2014). Epidemiology of Bluetongue in India. *Transboundary and Emerging diseases*.

Rao, S., Raju, D. & Suhasini, K. (2010). Shift in Buffalo Population and Composition in relation to Cattle during Post Green Revolution period in Andhra Pradesh, India. *Italian Journal of Animal Science,* ***6***(2s), 1394-1399.

Rasmusson, E. M. & Carpenter, T. H. (1983). The relationship between eastern equatorial Pacific sea surface temperatures and rainfall over India and Sri Lanka. *Monthly Weather Review,* ***111***(3), 517-528.

Ravishankar, C., Krishnan, N. G., Mini, M. & Jayaprakasan, V. (2005). Seroprevalence of bluetongue virus antibodies in sheep and goats in Kerala State, India. *Revue Scientifique et Technique (International Office of Epizootics)*, ***24***(3), 953-958.

Reddy, C. S. & Hafeez, M. (2008). Studies on certain aspects of prevalence of *Culicoides* species. *The Indian Journal of Animal Sciences*, ***78***(2).

Reddy, C. V. S. & Hafeez, M. (2008). Studies on certain aspects of prevalence of *Culicoides* species. *Indian Journal of Animal Sciences*, ***78***(2), 138-142.

Reuter, H. I., Nelson, A. & Jarvis, A. (2007). An evaluation of void-filling interpolation methods for SRTM data. *International Journal of Geographical Information Science*, ***21***(9), 983-1008.

Richardson, S., Abellan, J. J. & Best, N. (2006). Bayesian spatio-temporal analysis of joint patterns of male and female lung cancer risks in Yorkshire (UK). *Statistical Methods in Medical Research*, ***15***(4), 385-407.

Ripley, B. D. (2002). Modern applied statistics with S. Springer.

Robertson, C., Nelson, T. A., MacNab, Y. C. & Lawson, A. B. (2010). Review of methods for space–time disease surveillance. *Spatial and Spatio-temporal Epidemiology*, ***1***(2), 105-116.

Robinson, T. (2000). Spatial statistics and geographical information systems in epidemiology and public health. *Advances in Parasitology*, ***47***, 81-128.

Roger A. Pielke. (1997). *Hurricanes: Their nature and impacts on society*. John Wiley & Sons.

Rogers, D. (2006). Models for vectors and vector-borne diseases. *Advances in Parasitology, **62***, 1-35.

Rogers, D. J. & Packer, M. J. (1993). Vector-borne diseases, models, and global change. *The Lancet*, ***342***(8882), 1282-1284.

Rogers, D. J. & Randolph, S. E. (1993). Distribution of tsetse and ticks in Africa: past, present and future. *Parasitology Today,* ***9***(7), 266-271.

Rogers, D. J. & Randolph, S. E. (2003). Studying the global distribution of infectious diseases using GIS and RS. *Nature Reviews Microbiology*, ***1***(3), 231-237.

Rogers, D. J. (2015). Dengue: recent past and future threats. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, ***370***(1665), 20130562.

Rogers, D. J. Randolph, S. E., Snow, R. W. & Hay, S. I. (2002). Satellite imagery in the study and forecast of malaria. *Nature,* ***415***(6872), 710-715.

Rogers, D., Hay, S. & Packer, M. (1996). Predicting the distribution of tsetse flies in West Africa using temporal Fourier processed meteorological satellite data. *Annals of Tropical Medicine and Parasitology,* ***90***(3), 225-242.

Rosenblum, M. G., Pikovsky, A. S. & Kurths, J. (1996). Phase synchronization of chaotic oscillators. *Physical Review Letters*, ***76***(11), 1804.

Rue, H. & Held, L. (2005). *Gaussian Markov random fields: theory and applications*: CRC Press.

Rue, H., Martino, S. & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series b (Statistical Methodology), **71***(2), 319-392.

Ruiz, D., Poveda, G., Vélez, I. D., Quiñones, M. L., Rúa, G. L., Velásquez, L. E. & Zuluaga, J. S. (2006). Modelling entomological-climatic interactions of *Plasmodium falciparum* malaria transmission in two Colombian endemic-regions: contributions to a National Malaria Early Warning System. *Malaria Journal*, ***5***(1), 66.

Ruiz-Fons, F., Reyes-García, Á. R., Alcaide, V. & Gortázar, C. (2008). Spatial and temporal evolution of bluetongue virus in wild ruminants, Spain. *Emerging Infectious Diseases, **14***(6), 951.

Saegerman, C., Berkvens, D. & Mellor, P. S. (2008). Bluetongue epidemiology in the European Union. *Emerging Infectious Diseases, **14***(4), 539.

Saegerman, C., Hubaux, M., Urbain, B., Lengele, L. & Berkvens, D. (2007). Regulatory issues surrounding the temporary authorisation of animal vaccination in emergency situations: the example of bluetongue in Europe. *Revue sScientifique et Technique-Office International des épizooties, **26***(2), 395.

Sanders, C. J., Shortall, C. R., Gubbins, S., Burgin, L., Gloster, J., Harrington, R., Reynolds, D. R., Mellor, P. S. & Carpenter, S. (2011). Influence of season and meteorological parameters on flight activity of *Culicoides* biting midges. *Journal of Applied Ecology,* ***48***(6), 1355-1364.

Sapre, S. (1964). An outbreak of bluetongue in goats and sheep. *Vet. Rev,* ***15***, 78-80.

Scharlemann, J. P., Benz, D., Hay, S. I., Purse, B. V., Tatem, A. J., Wint, G. W. & Rogers, D. J. (2008). Global data for ecology and epidemiology: a novel algorithm for temporal Fourier processing MODIS data. *PLoS One*, ***3***(1), e1408.

Schrödle, B. & Held, L. (2011). A primer on disease mapping and ecological regression using INLA. *Computational Statistics,* ***26***(2), 241-258.

Schwartz-Cornil, I., Mertens, P.P., Contreras, V., Hemati, B., Pascale, F., Bréard, E., Mellor, P.S., MacLachlan, N.J. & Zientara, S. (2008). Bluetongue virus: virology, pathogenesis and immunity. *Veterinary Research*, ***39***(5), p.1.

Scrucca, L., Scrucca, M. L. & Imports, M. (2014). Package 'qcc'. In: http://cran.r-project.org/web/packages/qcc/qcc.pdf.

Searle, K., Blackwell, A., Falconer, D., Sullivan, M., Butler, A. & Purse, B. (2013). Identifying environmental drivers of insect phenology across space and time: *Culicoides* in Scotland as a case study. *Bulletin of Entomological Research,* ***103***(02), 155-170.

Sedda, L., Brown, H. E., Purse, B. V., Burgin, L., Gloster, J. & Rogers, D. J. (2012). A new algorithm quantifies the roles of wind and midge flight activity in the bluetongue epizootic in northwest Europe.

Proceedings of the Royal Society B: Biological Sciences, ***279***(1737), 2354-2362.

Sellers, R. & Mellor, P. (1993). Temperature and the persistence of viruses in *Culicoides* spp. during adverse conditions. *Revue Scientifique et Technique (International Office of Epizootics),* ***12***(3), 733-755.

Selvaraju, B., Rajendran, D., Kannan, D. & Geetha, M. (2013). Multiple linear regression model for forecasting Bluetongue disease outbreak in sheep of North-west agroclimatic zone of Tamil Nadu, India. *Veterinary World,* ***6***(6), 321-324.

Sen, P. & Gupta, S. D. (1959). Studies on Indian *Culicoides* (Ceratopogonidae: Diptera). *Annals of the Entomological Society of America,* ***52***(5), 617-630.

Sendow, I., Daniels, P., Cybinski, D., Young, P. & Ronohardjo, P. (1991). Antibodies against certain bluetongue and epizootic haemorrhagic disease viral serotypes in Indonesian ruminants. *Veterinary Microbiology,* ***28***(1), 111-118.

Sendow, I., Daniels, P., Cybinski, D., Young, P. & Ronohardjo, P. (1991). Antibodies against certain bluetongue and epizootic haemorrhagic disease viral serotypes in Indonesian ruminants. *Veterinary Microbiology*, ***28***(1), 111-118.

Shapiro, A. E., Tukahebwa, E. M., Kasten, J., Clarke, S. E., Magnussen, P., Olsen, A., Kabatereine, N. B., Ndyomugyenyi, R. & Brooker, S. (2005). Epidemiology of helminth infections and their relationship to clinical malaria in southwest Uganda. *Transactions of the Royal Society of Tropical Medicine and Hygiene,* ***99***(1), 18-24.

Shaw, A.E., Brüning-Richardson, A., Morrison, E.E., Bond, J., Simpson, J., Ross-Smith, N., Alpar, O., Mertens, P.P. & Monaghan, P. (2013). Bluetongue virus infection induces aberrant mitosis in mammalian cells.*Virology Journal,* ***10***(1), p.319.

Singh, N. & Sharma, V. (2002). Patterns of rainfall and malaria in Madhya Pradesh, central India. *Annals of Tropical Medicine and Parasitology,* ***96***(4), 349-359.

Singh, R., Saravanan, P., Sreenivasa, B., Singh, R. & Bandyopadhyay, S. (2004). Prevalence and distribution of peste des petits ruminants virus infection in small ruminants in India. *Revue Scientifique et Technique (International Office of Epizootics),* ***23***(3), 807-819.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* ***64***(4), 583-639.

Sreenivasulu, D., Subba, R. M., Reddy, Y. & Gard, G. (2003). Overview of bluetongue disease, viruses, vectors, surveillance and unique features: the Indian sub-continent and adjacent regions. *Veterinaria Italiana,* ***40***(3), 73-77.

Stein, M. L., Chi, Z. & Welty, L. J. (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* ***66***(2), 275-296.

Stenseth, N. C., Samia, N. I., Viljugrein, H., Kausrud, K. L., Begon, M., Davis, S., Leirs, H., Dubyanskiy, V., Esper, J. & Ageyev, V. S. (2006). Plague dynamics are driven by climate variation.

Proceedings of the National Academy of Sciences, ***103***(35), 13110-13115.

Sudhindra, Yathinder P.V. & Rajasekhar M. (1997). Haemorrhagic septicaemia and black quarter forecasting models for India. *Epidemic!. sante anim*., 1997, 31-32.

Sun, Y., Li, B. & Genton, M. G. (2012). Geostatistics for large datasets *Advances and challenges in space-time modelling of natural events* (pp. 55-77): Springer.

Tabachnick, W. (2003). *Culicoides* and the global epidemiology of bluetongue virus infection. *Veterinaria Italiana,* ***40***(3), 144-150.

Tatem, A., Baylis, M., Mellor, P., Purse, B., Capela, R., Pena, I. & Rogers, D. (2003). Prediction of bluetongue vector distribution in Europe and north Africa using satellite imagery. *Veterinary Microbiology,* ***97***(1), 13-29.

Tatsuoka, M. M. & Lohnes, P. R. (1988). *Multivariate analysis: Techniques for educational and psychological research*. Macmillan Publishing Co, Inc.

Taylor, W. (1986). The epidemiology of bluetongue. *Revue Scientifique et Technique de l'Office international des Epizooties,* **5**, 351-356.

Thomson, M. C., Doblas-Reyes, F. J., Mason, S. J., Hagedorn, R., Connor, S. J., Phindela, T. & Palmer, T. N. (2006). Malaria early warnings based on seasonal climate forecasts from multi-model ensembles. *Nature*, ***439*** (7076), 576-579.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine, **16***(4), 385-395.

Trenberth, K. E. (1997). The definition of el nino. *Bulletin of the American Meteorological Society, **78***(12), 2771-2777.

Udupa, G. K. (2001). *Culicoides Spp.(Diptera: Ceratopogonidae) Associated With Livestock And Their Relevance To Bluetongue Infection In Tamil Nadu.* PhD thesis submitted to Tamil Nadu Veterinary and Animal Sciences University.

Vecchia, A. V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 297-312.

Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with S*: Springer.

Verwoerd, D. & Erasmus, B. (2004). Bluetongue. *Infectious diseases of livestock, **2***, 1201-1220.

Wade, T. J., Calderon, R. L., Brenner, K. P., Sams, E., Beach, M., Haugland, R., Wymer, L. & Dufour, A. P. (2008). High sensitivity of children to swimming-associated gastrointestinal illness: results using a rapid assay of recreational water quality. *Epidemiology, **19***(3), 375-383.

Walker, A. R. (1977). Seasonal fluctuations of Culicoides species (Diptera: Ceratopogonidae) in Kenya. *Bulletin of Entomological Research, **67***(2), 217-233.

Waller, L. A., Goodwin, B. J., Wilson, M. L., Ostfeld, R. S., Marshall, S. L. & Hayes, E. B. (2007). Spatio-temporal patterns in county-level incidence and reporting of Lyme disease in the northeastern United States, 1990–2000. *Environmental and Ecological Statistics*, ***14***(1), 83-100.

Walter, S. & Tiemeier, H. (2009). Variable selection: current practice in epidemiological studies. *European Journal of Epidemiology,* ***24***(12), 733-736.

Wangdi, K., Singhasivanon, P., Silawan, T., Lawpoolsri, S., White, N. J. & Kaewkungwal, J. (2010). Development of temporal modelling for forecasting and prediction of malaria infections using time-series and ARIMAX analyses: a case study in endemic districts of Bhutan. *Malaria Journal*, ***9***(251), 251-259.

Ward, M. P. & Carpenter, T. E. (2000). Analysis of time–space clustering in veterinary epidemiology. *Preventive Veterinary Medicine*, ***43***(4), 225-237.

Wellby, M. P., Baylis, M., Rawlings, P. & Mellor, P. S. (1996). Effect of temperature on survival and rate of virogenesis of African horse sickness virus in *Culicoides variipennis sonorensis* (Diptera: Ceratopogonidae) and its significance in relation to the epidemiology of the disease. *Bulletin of Entomological Research*, ***86***(06), 715-720.

Welte, V. R. & Teran, M. V. (2004). Emergency Prevention System (EMPRES) for Transboundary Animal and Plant Pests and Diseases. The EMPRES-Livestock: An FAO Initiative. *Annals of the New York Academy of Sciences*, ***1026***(1), 19-31.

Wikle, C. K. (2002). Spatial modeling of count data: A case study in modelling breeding bird survey data on large spatial domains. *Chapman and Hall*, 199-209.

Wilson, A. J. & Mellor, P. S. (2009). Bluetongue in Europe: past, present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences, **364***(1530), 2669-2681.

Wilson, A. J., Ribeiro, R. & Boinas, F. (2013). Use of a Bayesian network model to identify factors associated with the presence of the tick *Ornithodoros erraticus* on pig farms in southern Portugal. *Preventive Veterinary Medicine, **110***(1), 45-53.

Wirth, W. W. & Hubert, A. A. (1989). *The Culicoides of Southeast Asia (Diptera: Ceratopogonidae).*

Wittmann, E., Mellor, P. & Baylis, M. (2001). Using climate data to map the potential distribution of *Culicoides imicola* (Diptera: Ceratopogonidae) in Europe. *Revue Scientifique et Technique-Office International des Epizooties*, ***20***(3), 731-736.

Wittmann, E., Mellor, P. & Baylis, M. (2002). Effect of temperature on the transmission of orbiviruses by the biting midge, *Culicoides sonorensis*. *Medical and Veterinary Entomology*, ***16***(2), 147-156.

World Bank. (2001). *World Development Report 2000-2001: Attacking Poverty*. El Banco.

Xie, P., Yarosh, Y., Love, T., Janowiak, J. E. & Arkin, P. A. (2002). *A real-time daily precipitation analysis over South Asia.* Paper presented at the Preprints of the 16th Conference of Hydrology: 2002; Orlando.

Yanase, T., Matsumoto, Y., Matsumori, Y., Aizawa, M., Hirata, M., Kato, T., Shirafuji, H., Yamakawa, M., Tsuda, T. & Noda, H. (2013). Molecular identification of field-collected *Culicoides* larvae in the southern part of Japan. *Journal of Medical Entomology,* ***50***(5), 1105-1110.

Zhang, Y., Bi, P. & Hiller, J. E. (2008). Climate change and the transmission of vector-borne diseases: a review. *Asia-Pacific Journal of Public Health,* ***20***(1), 64-76.

Zhang, Y., Bi, P. & Hiller, J. E. (2010). Meteorological variables and malaria in a Chinese temperate city: A twenty-year time-series data analysis. *Environment International,* ***36***(5), 439-445.

Zhou, G., Minakawa, N., Githeko, A. K. & Yan, G. (2004). Association between climate variability and malaria epidemics in the East African highlands. *Proceedings of the National Academy of Sciences of the United States of America,* ***101***(8), 2375-2380.

Zubair, L. & Ropelewski, C. F. (2006). The strengthening relationship between ENSO and northeast monsoon rainfall over Sri Lanka and southern India. *Journal of Climate,* ***19***(8), 1567-1575.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A. & Smith, G. M. (2009). Zero-truncated and zero-inflated models for count data *Mixed effects models and extensions in ecology with R* (pp. 261-293): Springer.