Mapping and understanding the distributions of potential vector mosquitoes in the UK: New methods and applications

Nicholas Golding

Linacre College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2013

Mapping and understanding the distributions of potential vector mosquitoes in the UK: New methods and applications

Nicholas Golding

Linacre College
University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Trinity 2013

A number of emerging vector-borne diseases have the potential to be transmitted in the UK by native mosquitoes. Human infection by some of these diseases requires the presence of communities of multiple vector mosquito species. Mitigating the risk posed by these diseases requires an understanding of the spatial distributions of the UK mosquito fauna. Little empirical data is available from which to determine the distributions of mosquito species in the UK. Identifying areas at risk from mosquito-borne disease therefore requires statistical modelling to investigate and predict mosquito distributions.

This thesis investigates the distributions of potential vector mosquitoes in the UK at landscape to national scales. A number of new methodological approaches for species distribution modelling are developed. These methods are then used to map and understand the distributions of mosquito communities with the potential to transmit diseases to humans.

Chapter 2 reports the establishment of substantial populations of the West Nile virus (WNV) vector mosquito *Culex modestus* in wetlands in southern England. This represents a drastic shift in the species' known range and an increase in the risk of WNV transmission where *Cx. modestus* is present.

Chapter 3 develops and applies a new species interaction distribution model which identifies fish and ditch shrimp of the genus *Palaemonetes* as predators which may restrict the distribution of the potential WNV vector community in these wetlands.

Chapter 4 develops a number of methods to make robust predictions of the probability of presence of a species from presence-only data, by eliciting and applying estimates of the species' prevalence.

Chapter 5 introduces a new Bayesian species distribution modelling approach which outperforms existing methods and has number of useful features for dealing with poor-quality data.

Chapter 6 applies methods developed in the previous two chapters to produce the first high-resolution distribution maps of potential vector mosquitoes in the UK. These maps identify several wetland areas where vector communities exist which could maintain WNV transmission in birds and transmit it to humans.

This thesis makes significant contributions to our understanding of the distributions of UK mosquito species. It also provides methods for species distribution modelling which could be widely applied in ecology and epidemiology.

Acknowledgements

I would like to thank my supervisors David Rogers, Beth Purse and Miles Nunn for all of the support and freedom they have given me over the past three and a half years. Without their sage advice, technical help and relentless enthusiasm this thesis may never have happened and would certainly have been a lot less fun.

Thanks also to Steffi Schäfer, David Benz, David Morley, Luigi Sedda and Nell Godfrey for their advice, help and for all the fruitful and fruitless discussions.

The fieldwork portion of this thesis would not have been possible without the permission and help of various reserve managers and land owners, though special thanks go to Steve Gordon at Elmley NNR on the Isle of Sheppey. Thanks also go to Gay Gibson and Simon Springate at NRI, University of Greenwich for all their help, to Mick Peacey, Rachel Madison, John Day and Tom August at the Centre for Ecology & Hydrology for their good company in the field and to two farmers of the Somerset levels for their patience and tow rope.

I would like to thank all of my friends and family for their love and support over the last few years and particularly to Holly, who has supported me more than she knows.

The work in this thesis was supported by funding from the Centre for Ecology & Hydrology.

"*C. pipiens* is so common and widespread in Britain that a compilation of its locality records would be of no scientific interest."

John. F. Marshall in The British Mosquitoes

Contents

Chapter 1

General introduction

Emerging vector-borne diseases

The last two decades have seen dramatic expansions in the global distributions of a number of vector-borne diseases. These include the introduction and rapid spread of West Nile virus (WNV) in North America (Hayes, 2001) and Bluetongue virus (BTV) into southern and then northern Europe (Purse *et al.*, 2005; Carpenter *et al.*, 2009) as well as outbreaks of chikungunya fever in Italy and dengue fever in France (ECDC, 2007; La Ruche *et al.*, 2010). The spread of these diseases has resulted in human morbidity and mortality and severe economic losses (Wilson & Mellor, 2009). Understanding the factors responsible for these changing distributions and predicting future spread has become an important issue in global health research (Kilpatrick & Randolph, 2012).

In many cases, spread of a pathogen to a new region follows shortly after the introduction of an efficient vector (Reiter & Sprenger, 1987). Drastic increases in the volume of global traffic in recent history have been identified as a major driver of range expansions of vectors and the diseases they transmit (Tatem *et al.*, 2006). In other cases, such as the establishment of WNV in North America and BTV in northern Europe, suitable vectors are already present in the region (Randolph & Rogers, 2010). Explaining range shifts in these pathogens is less straightforward, though climate change (Purse *et al.*, 2005; Guis *et al.*, 2012), land use change (Lambin *et al.*, 2010), socio economic factors (Randolph, 2010) and changes in disease control (Hay *et al.*, 2002) are all likely to play a role.

Whilst the complexity of these situations makes it almost impossible to predict the arrival of specific diseases, the identification of areas at risk of disease establishment and spread has received a great deal of focus (Medlock *et al.*, 2007b; Hartemink *et al.*, 2009; Lindsay *et al.*, 2010).

Spatial variation in disease risk

The ecology of vector-borne pathogens is complex. To persist they require not only suitable environmental conditions for their own biological processes, but also the presence and sufficient abundance of susceptible host and vector species. The spatial distributions of such diseases are restricted to where all of these components coincide (Reisen, 2010). As a result risk of vector-borne diseases tends to be spatially heterogeneous (Bousema *et al.*, 2012).

Identifying regions at risk from vector-borne diseases therefore requires an understanding of how the ecology of pathogens, hosts and vectors drives their joint distributions (LoGiudice *et al.*, 2003; Beketov *et al.*, 2010; Ferguson *et al.*, 2010) and interactions.

Mosquito-borne disease risk in the UK

At present, no mosquito-borne diseases are thought to be transmitted to humans in the UK (Medlock *et al.*, 2007b). However several mosquito species present in the UK have the potential to act as vectors of human pathogens and pose a risk to public health (Ramsdale & Gunn, 2005; Medlock *et al.*, 2005).

Malaria

Human malaria was previously transmitted in parts of the UK and may have been responsible for significant morbidity and mortality (Dobson (1980), though see Hutchinson & Lindsay (2006a)). It is unclear which of the human-infecting species of *Plasmodia* were the responsible pathogens, though the salt-marsh mosquito *Anopheles atroparvus* is widely considered to have been the vector (Dobson, 1980). Whilst much of the UK may have climatic conditions suitable for the transmission of *Plasmodium vivax* (Lindsay & Thomas, 2001; Lindsay *et al.*, 2010), transmission could only occur where the anopheline vectors feed on humans at a relatively high rate. Current and future risk of malaria is likely to be much lower than in the past, since the condi-

tions which led to high levels of human biting by *An. atroparvus* no longer occur (Bruce-Chwatt & De Zulueta, 1980; Kuhn *et al.*, 2003). Any transmission of malaria would therefore be restricted to areas where human hosts coincide with abundant populations of suitable anopheline vectors remote from preferred vertebrate hosts of these vectors.

Arboviruses

The aedine mosquito species *Aedes aegypti* and *Ae. albopictus* are efficient vectors of established and emerging mosquito-borne viruses of significant public health importance (Reiter, 2010b) but are currently absent from the UK. However a number of other arboviral infections of humans could be transmitted by native mosquitoes, including WNV as well as the Sindbis (SV) and Tahyna (TV) viruses (Medlock *et al.*, 2007b) which are present in Europe.

WNV is responsible for periodic epidemics and epizootics in Europe, particularly in the wetlands in the south of the continent (Pradier *et al.*, 2008). During an outbreak of the disease in Greece in 2010, 262 human cases were reported, resulting in 197 cases of neuroinvasive disease and 33 deaths (Danis *et al.*, 2011). SV and TV are thought to produce milder symptoms and no fatalities have been reported from either disease (Medlock *et al.*, 2007b). However the difficulty in identifying the causes of viral encephalitis cases make it problematic to establish their true burden (Davison *et al.*, 2003).

Neutralizing antibodies to both WNV and SV have been detected in wild birds (Buckley *et al.*, 2003) and sentinel chickens (Buckley *et al.*, 2006) in the UK. Whilst these findings suggest that enzootic cycles of these viruses may already be present, confirmation of this by virus isolation has not so far been achieved.

These three pathogens share similar ecology, being maintained in an enzootic cycle in birds by many of the same mosquito species (Lundström, 1994). Humans and other mammals are essentially 'dead-end' hosts for these viruses, developing an insufficient level of viraemia to reinfect mosquitoes (Bunning & Bowen, 2001).

Transmission to humans therefore requires both enzootic vectors which habitually bite birds and sustain the enzootic cycle, and 'bridge vectors' which will bite both birds and mammals, transmitting the viruses to susceptible human hosts (Reiter, 2010a).

Other public health concerns

Dirofilariasis (caused by the parasitic nematodes *Dirofilaria immitis* and *D. repens*) is predominantly a disease of dogs, but is considered an emerging zoonotic infection in southern Europe (Pampiglione *et al.*, 2001; Genchi *et al.*, 2005). The parasites can be vectored by a wide range of common mosquito species (Svobodova & Misonova, 2005) and environmental conditions in southern and central England appear to be suitable for extrinsic incubation (Medlock *et al.*, 2007a). However human cases are rare and typically result in only relatively mild pathology.

By far the most immediate public health consequence of UK mosquitoes is nuisance biting. Whilst mosquito bites can occur almost anywhere in the UK during the summer months, major biting nuisance tends to highly localized in areas of large populations of species such as *Ochlerotatus detritus* (Hutchinson & Lindsay, 2006b). The Kent and Essex coasts have been the site of significant problems, with mosquito biting rates of 200 per hour recorded (Ramsdale & Snow, 1995). Whilst mosquito bites are a mild irritation in most cases, severe allergic reactions are not uncommon (Malcolm, 2009).

Mosquito distributions

Most potential vector mosquitoes in the UK are rare by comparison with susceptible human and avian hosts. Spatial variation in the risk of these diseases is therefore particularly dependent on the distributions of suitable vector mosquitoes. Consequently, identification of areas at risk from mosquito-borne disease depends on knowledge of the distributions of potential vectors (Medlock *et al.*, 2005). Mapping these distributions and understanding what drives them is essential to predicting and mitigating

risk from mosquito-borne disease.

Empirical data on the distributions of mosquito species in the UK comprise a limited number of occurrence records. In order to map spatial variation in mosquito-borne disease risk we therefore need to construct predictive models of their distributions. To accurately predict these distributions at a national scale requires methods to account for the various sources of bias in the poor-quality occurrence data available.

Species distribution modelling

Species distribution models (SDMs) are a range of quantitative models of the spatial distributions of one or more species. The ability of these models to produce continuous maps of predicted distributions has led to their wide use in ecology and related fields (Elith & Leathwick, 2009). Among other applications, SDMs have been used to investigate ecological and evolutionary theories (Anderson *et al.*, 2002; Hugall *et al.*, 2002), to aid wildlife conservation by identifying critical habitats for endangered species (Elith & Leathwick, 2006) and even to detect new species (Raxworthy *et al.*, 2003). SDMs have also become widely used in public health to map the current and potential distributions of diseases and their vectors (Rogers, 2006; Bhatt *et al.*, 2013) as well as incriminating potential hosts and vectors of emerging diseases (Peterson *et al.*, 2004; Purse *et al.*, 2007) and identifying complex inter-species relationships in disease ecosystems (Stephens *et al.*, 2009).

Extensions

The most widely used approach to distribution modelling is to relate the presence or absence of individuals of a species at a set of sites to environmental variables measured at those sites. These models have been referred to as 'niche' models owing to their basis in the Grinellian and Hutchinsonian concepts of niche (Grinnell, 1904; Hutchinson, 1957; Warren, 2012).

In recent years however, a number of extensions to these SDMs have been developed (Guisan & Thuiller, 2005). Spatially explicit SDMs have been proposed which account for non-environmental spatial correlation in distributions arising from population processes (Gelfand *et al.*, 2006; Kearney & Porter, 2009). Mechanistic, process-based models have incorporated additional aspects of ecological theory into SDMs (Dormann *et al.*, 2012; García-Valdés *et al.*, 2013). Species-interaction distribution models have been developed which attempt to describe the distributions of entire ecological communities by modelling biotic interactions between species (Ovaskainen

et al., 2010; Kissling *et al.*, 2011; Wisz *et al.*, 2013).

These extensions to the widely used SDM concept have great potential to expand our understanding of ecology and identify the various drivers of species' distributions. Whilst these approaches have so far been little used by epidemiologists (though spatial models are increasingly used, see e.g. Gething *et al.* (2006); Linard (2009)) they have similar potential for public health research.

Practical issues

Whilst investigating the drivers of species' distributions is a fundamental question in both ecology and epidemiology, production of accurate distribution maps to guide policy decisions is often the immediate aim. In many cases the only data available from which to construct distribution maps are scarce and of variable quality, precluding the use of some of the more complex models described above. Consequently, the more flexible and robust correlative SDM approaches are more applicable in many cases and methods for dealing with the various sources of bias in the available data are crucial.

Presence-only data

Often, the only available empirical information on the distribution of a species consists of sites where the species has been recorded (Graham *et al.*, 2004). Since these presence-only data are far from the randomly-sampled presence-absence data required to fit most statistical models to binary data, different approaches are needed. A common approach is to augment occurrence data with randomly-sampled background data (also referred to as pseudo-absence data) and apply a standard presence-absence statistical model (Barbet-Massin *et al.*, 2012). Unfortunately this approach violates several assumptions of presence-absence statistical models and can lead to bias in model predictions (Ward, 2007). In general, distribution maps generated using these methods do not represent the probability that the species is present.

Recently, models have been proposed to generate predictions of probability of

presence using presence-background data (Ward *et al.*, 2009; Phillips & Elith, 2011). These approaches have yet to see widespread use, due to their novelty and the requirement for an estimate of the species' prevalence which can be hard to acquire (Phillips & Elith, 2013).

Other forms of bias

Occurrence data used to fit SDMs is often collated from a variety of different sources. These may include museum collections (Graham *et al.*, 2004), records provided by members of the public (Dickinson *et al.*, 2012) or published scientific literature. Records resulting from these sources are typically of variable quality and subject to various forms of bias.

In many cases the occurrence data has spatial precision lower than that of the gridded environmental data used to fit the model (Mcpherson *et al.*, 2006). If this is not accounted for, it can lead to regression dilution in model predictions (Frost & Thompson, 2000; McInerny *et al.*, 2011). Such occurrence data tend to be subject to significant spatial bias in recording effort - individuals are more likely to be recorded from some areas than others. This can bias model predictions towards the environmental signature of recording effort, rather than of the species being considered (Phillips *et al.*, 2009).

Given the paucity of the data and potential implications of the resulting predictions, a robust treatment of the uncertainties associated with such distribution maps is desirable, though rarely implemented (Elith *et al.*, 2002).

Aims and structure of the thesis

The aim of this thesis is to add to our understanding of the spatial distributions of potential vector mosquitoes in the UK. To achieve this aim using the limited empirical data available, there is a strong focus on the development of new methods and approaches for modelling species' distributions.

Structure

Chapter 2 reports the presence in southern England of substantial populations of *Culex modestus*, an efficient vector of WNV not previously thought to be present in the UK. The possible origins of these populations and the implications of its presence on disease risk are discussed. In Chapter 3, a novel species interaction distribution model is applied to larval survey data from wetland sites where populations of *Cx. modestus* are established. This allows us to identify biotic interactions such as competition and predation which may influence the distribution of the community of potential WNV vector mosquitoes in these locations.

The rest of the thesis then focuses on applying SDMs to map and understand the national-scale distributions of potential vector mosquitoes in the UK. Since the only available data on broad-scale mosquito distributions in the UK consist of a limited number of presence-only records, a number of new methods are developed to deal with these issues. Chapter 4 develops a method to elicit accurate estimates of species' prevalence from expert opinion and approaches to apply these estimates to produce robust predictions of probability of presence from presence-only data. Chapter 5 introduces GRaF, a new Bayesian approach to SDM. A variety of useful features for dealing with poor-quality data are demonstrated and the method is shown to have higher predictive accuracy than existing models.

Chapter 6 combines the methods developed in Chapters 5 and 6 and applies them to generate high resolution predicted distribution maps for potential vector mosquitoes in the UK from presence only distribution data. These distribution maps

are combined to map the distributions of communities of potential avian arbovirus vectors, identifying areas of greatest risk in the case of introduction of WNV, SV or TV. Land cover correlates of the distributions of individual species and communities are explored.

Finally, Chapter 7 provides a general discussion of the thesis and its implications for species distribution modelling and future risk of mosquito-borne pathogens in the UK.

Format

Chapters 2 to 6 have been prepared in the format of scientific journal articles, rather than traditional thesis chapters. Chapter 2 has been published in the journal *Parasites & Vectors* whilst the rest are yet to be published. The significant majority of each chapter is my own work and author contributions are detailed on the title page of each chapter.

References

Anderson, R.P., Peterson, A. & Gómez-Laverde, M. (2002) Using niche-based GIS modeling to test geographic predictions of competitive exclusion and competitive release in South American pocket mice. *Oikos*, **1**, 3–16.

Barbet-Massin, M., Jiguet, F., Albert, C.H. & Thuiller, W. (2012) Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution*, **3**, 327–338.

Beketov, M.A., Yurchenko, Y.A., Belevich, O.E. & Liess, M. (2010) What Environmental Factors are Important Determinants of Structure, Species Richness, and Abundance of Mosquito Assemblages? *Journal of Medical Entomology*, **47**, 129–139.

Bhatt, S., Gething, P.W., Brady, O.J., Messina, J.P., Farlow, A.W., Moyes, C.L., Drake, J.M., Brownstein, J.S., Hoen, A.G., Sankoh, O., Myers, M.F., George, D.B., Jaenisch, T., Wint, G.R.W., Simmons, C.P., Scott, T.W., Farrar, J.J. & Hay, S.I. (2013) The global distribution and burden of dengue. *Nature*.

Bousema, T., Griffin, J.T., Sauerwein, R.W., Smith, D.L., Churcher, T.S., Takken, W., Ghani, A.C., Drakeley, C. & Gosling, R.D. (2012) Hitting Hotspots: Spatial Targeting of Malaria for Control and Elimination. *PLoS Medicine*, **9**, e1001165.

Bruce-Chwatt, L. & De Zulueta, J. (1980) *The Rise and Fall of Malaria in Europe: A Historico-epidemiological Study*. Oxford University Press.

Buckley, A., Dawson, A., Moss, S.R., Hinsley, S.A., Bellamy, P.E. & Gould, E.A. (2003) Serological evidence of West Nile virus, Usutu virus and Sindbis virus infection of birds in the UK. *Journal of General Virology*, **84**, 2807–2817.

Buckley, A., Gould, E.A. & Dawson, A. (2006) Detection of seroconversion to West Nile virus, Usutu virus and Sindbis virus in UK sentinel chickens. *Virology Journal*, **3**, 71.

Bunning, M. & Bowen, R. (2001) Experimental infection of horses with West Nile virus and their potential to infect mosquitoes and serve as amplifying hosts. *Annals of the New York Academy of Sciences*, **951**, 338–339.

Carpenter, S.T., Wilson, A.J. & Mellor, P.S. (2009) Culicoides and the emergence of bluetongue virus in northern Europe. *Trends in Microbiology*, **17**, 172–8.

Danis, K., Papa, A., Papanikolaou, E., Dougas, G., Terzaki, I., Baka, A., Vrioni, G., Kapsimali, V., Tsakris, A., Kansouzidou, A., Tsiodras, S., Vakalis, N., Bonovas, S. & Kremastinou, J. (2011) Ongoing outbreak of West Nile virus infection in humans, Greece, July to August 2011. *Euro Surveillance*, **16**, 1–5.

Davison, K., Crowcroft, N., Ramsay, M., Brown, D. & Andrews, N. (2003) Viral encephalitis in England, 19891998: What did we miss? *Emerging Infectious Diseases*, **9**, 234–240.

Dickinson, J.L., Shirk, J., Bonter, D., Bonney, R., Crain, R.L., Martin, J., Phillips, T. & Purcell, K. (2012) The current state of citizen science as a tool for ecological research and public engagement. *Frontiers in Ecology and the Environment*, **10**, 291–297.

Dobson, M. (1980) Marsh fever -The geography of malaria in England. *Journal of Historical Geography*, **6**, 357–389.

Dormann, C.F., Schymanski, S.J., Cabral, J., Chuine, I., Graham, C.H., Hartig, F., Kearney, M., Morin, X., Römermann, C., Schröder, B. & Singer, A. (2012) Correlation and process in species distribution models: bridging a dichotomy. *Journal of Biogeography*, **39**, 2119–2131.

ECDC (2007) Mission Report: Chikungunya in Italy. Joint ECDC/WHO visit for a European risk assessment. Technical report.

Elith, J., Burgman, M.a. & Regan, H.M. (2002) Mapping epistemic uncertainties and

vague concepts in predictions of species distribution. *Ecological Modelling*, **157**, 313–329.

Elith, J. & Leathwick, J.R. (2006) Conservation prioritisation using species distribution modelling. A. Moilanen, K. Wilson & H. Possingham, eds., *Spatial Conservation Prioritization: Quantitative Methods and Computational Tools*, pp. 70–93. Oxford University Press, Oxford.

Elith, J. & Leathwick, J.R. (2009) Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 677–697.

Ferguson, H.M., Dornhaus, A., Beeche, A., Borgemeister, C., Gottlieb, M., Mulla, M.S., Gimnig, J.E., Fish, D. & Killeen, G.F. (2010) Ecology: A Prerequisite for Malaria Elimination and Eradication. *PLoS Medicine*, **7**, e1000303.

Frost, C. & Thompson, S.G. (2000) Correcting for regression dilution bias: comparison of methods for a single predictor variable. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **163**, 173–189.

García-Valdés, R., Zavala, M., Araújo, M.B. & Purves, D.W. (2013) Chasing a moving target: projecting climate changeinduced shifts in nonequilibrial tree species distributions. *Journal of Ecology*, **101**, 441–453.

Gelfand, A., Silander, J.A. & Wu, S. (2006) Explaining species distribution patterns through hierarchical modeling. *Bayesian Analysis*, **1**, 41–92.

Genchi, C., Rinaldi, L., Cascone, C., Mortarino, M. & Cringoli, G. (2005) Is heartworm disease really spreading in Europe? *Veterinary parasitology*, **133**, 137–48.

Gething, P.W., Noor, A.M., Gikandi, P.W., Ogara, E.A.A., Hay, S.I., Nixon, M.S., Snow, R.W. & Atkinson, P.M. (2006) Improving imperfect data from health management information systems in Africa using space-time geostatistics. *PLoS Medicine*, **3**, e271.

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A. (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution*, **19**, 497–503.

Grinnell, J. (1904) The origin and distribution of the chest-nut-backed chickadee. *The Auk*, **21**, 364–382.

Guis, H., Caminade, C., Calvete, C., Morse, A.P., Tran, A. & Baylis, M. (2012) Modelling the effects of past and future climate on the risk of bluetongue emergence in Europe. *Journal of the Royal Society, Interface*, **9**, 339–50.

Guisan, A. & Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8**, 993–1009.

Hartemink, N.A., Purse, B.V., Meiswinkel, R., Brown, H.E., de Koeijer, A., Elbers, a.R.W., Boender, G.J., Rogers, D.J. & Heesterbeek, J.a.P. (2009) Mapping the basic reproduction number (R) for vector-borne diseases: a case study on bluetongue virus. *Epidemics*, **1**, 153–61.

Hay, S.I., Rogers, D.J., Randolph, S.E., Stern, D.I., Cox, J., Shanks, G.D. & Snow, R.W. (2002) Hot topic or hot air? Climate change and malaria resurgence in East African highlands. *Trends in Parasitology*, **18**, 530–4.

Hayes, C.G. (2001) West Nile virus: Uganda, 1937, to New York City, 1999. *Annals of the New York Academy of Sciences*, **951**, 25–37.

Hugall, A., Moritz, C., Moussalli, A. & Stanisic, J. (2002) Reconciling paleodistribution models and comparative phylogeography in the Wet Tropics rainforest land snail Gnarosophia bellendenkerensis (Brazier 1875). *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 6112–7.

Hutchinson, G. (1957) Concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, **22**, 415 –427.

Hutchinson, R.A. & Lindsay, S.W. (2006a) Malaria and deaths in the English marshes. *Lancet*, **367**, 1947–1951.

Hutchinson, R.A. & Lindsay, S.W. (2006b) Perceived nuisance of mosquitoes on the Isle of Sheppey, Kent, UK. *Journal of Biosocial Science*, **38**, 707–12.

Kearney, M. & Porter, W. (2009) Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology Letters*, **12**, 334–50.

Kilpatrick, A.M. & Randolph, S.E. (2012) Drivers, dynamics, and control of emerging vector-borne zoonotic diseases. *Lancet*, **380**, 1946–55.

Kissling, W.D., Dormann, C.F., Groeneveld, J., Hickler, T., Kühn, I., McInerny, G.J., Montoya, J.M., Römermann, C., Schiffers, K., Schurr, F.M., Singer, A., Svenning, J.C., Zimmermann, N.E. & OHara, R.B. (2011) Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *Journal of Biogeography*, **39**, 2163–2178.

Kuhn, K.G., Campbell-Lendrum, D.H., Armstrong, B. & Davies, C.R. (2003) Malaria in Britain: Past, present, and future. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 9997–10001.

La Ruche, G., Souarès, Y., Armengaud, A., Peloux-Petiot, F., Delaunay, P., Desprès, P., Lenglet, A., Jourdain, F., Leparc-Goffart, I., Charlet, F., Ollier, L., Mantey, K., Mollet, T., Fournier, J.P., Torrents, R., Leitmeyer, K., Hilairet, P., Zeller, H.G., Van Bortel, W., Dejour-Salamanca, D., Grandadam, M. & Gastellu-Etchegorry, M. (2010) First two autochthonous dengue virus infections in metropolitan France, September 2010. *Euro Surveillance*, **15**, 19676.

Lambin, E.F., Tran, A., Vanwambeke, S.O., Linard, C. & Soti, V. (2010) Pathogenic landscapes: interactions between land, people, disease vectors, and their animal hosts. *International Journal of Health Geographics*, **9**, 54.

Linard, C. (2009) *Spatial and integrated modelling of the transmission of vector-borne and zoonotic infections*. Ph.D. thesis, Université catholique de Louvain.

Lindsay, S.W. & Thomas, C.J. (2001) Global warming and risk of vivax malaria in Great Britain. *Global Change and Human Health*, **2**, 80–84.

Lindsay, S.W., Hole, D.G., Hutchinson, R.a., Richards, S.a. & Willis, S.G.S.G. (2010) Assessing the future threat from vivax malaria in the United Kingdom using two markedly different modelling approaches. *Malaria Journal*, **9**, 70.

LoGiudice, K., Ostfeld, R.S., Schmidt, K.A. & Keesing, F. (2003) The ecology of infectious disease: Effects of host diversity and community composition on Lyme disease risk. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 567–571.

Lundström, J.O. (1994) Vector competence of Western European mosquitoes for arboviruses: A review of field and experimental studies. *Bulletin of the Society of Vector Ecologists*, **19**, 23–36.

Malcolm, C.A. (2009) Public health issues posed by mosquitoes An independent report. Technical report.

McInerny, G.J., Purves, D.W. & McIntyre, K.M. (2011) Fine-scale environmental variation in species distribution modelling : regression dilution , latent variables and neighbourly advice. *Methods in Ecology and Evolution*, **2**, 248–257.

Mcpherson, J.M., Jetz, W. & Rogers, D.J. (2006) Using coarse-grained occurrence data to predict species distributions at finer spatial resolutionspossibilities and limitations. *Ecological Modelling*, **192**, 499–522.

Medlock, J.M., Barrass, I., Taylor, M.A., Kerrod, E. & Leach, S. (2007a) Analysis of climatic predictions for extrinsic incubation of Dirofilaria in the United kingdom. *Vector-Borne and Zoonotic Diseases*, **7**, 4–14.

Medlock, J.M., Leach, S. & Snow, K.R. (2005) Potential transmission of West Nile virus in the British Isles: an ecological review of candidate mosquito bridge vectors. *Medical and Veterinary Entomology*, **19**, 2–21.

Medlock, J.M., Leach, S. & Snow, K.R. (2007b) Possible ecology and epidemiology of medically important mosquito-borne arboviruses in Great Britain. *Epidemiology and Infection*, **135**, 466–482.

Ovaskainen, O., Hottola, J. & Siitonen, J. (2010) Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology*, **91**, 2514–21.

Pampiglione, S., Rivasi, F., Angeli, G., Boldorini, R., Incensati, R.M., Pastormerlo, M., Pavesi, M. & Ramponi, a. (2001) Dirofilariasis due to Dirofilaria repens in Italy, an emergent zoonosis: report of 60 new cases. *Histopathology*, **38**, 344–54.

Peterson, A., Carroll, D.S., Mills, J.N. & Johnson, K.M. (2004) Potential mammalian filovirus reservoirs. *Emerging infectious diseases*, **10**, 2073–81.

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J.R. & Ferrier, S. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, **19**, 181–97.

Phillips, S.J. & Elith, J. (2011) Logistic methods for resource selection functions and presence-only species distribution models. *AAAI (Association for the Advancement of Artificial Intelligence*, pp. 1384–1389.

Phillips, S.J. & Elith, J. (2013) On Estimating Probability of Presence from Use-Availability or Presence-Background Data. *Ecology*.

Pradier, S., Leblond, A. & Durand, B. (2008) Land cover, landscape structure, and West Nile virus circulation in southern France. *Vector-Borne and Zoonotic Diseases*, **8**, 253–263.

Purse, B.V., Mccormick, B.J.J., Mellor, P.S., Baylis, M., Boorman, J.P.T., Borras, D., Burgu, I., Capela, R., Caracappa, S., Collantes, F., De Liberato, C., Delgado, J.A., Denison, E., Georgiev, G., Harak, M.E., de La Rocque, S., Lhor, Y., Lucientes, J., Mangana, O., Miranda, M.A., Nedelchev, N., Nomikou, K., Ozkul, A., Patakakis, M., Pena, I., Scaramozzino, P., Torina, A. & Rogers, D.J. (2007) Incriminating bluetongue virus vectors with climate envelope models. *Journal of Applied Ecology*, **44**, 1231–1242.

Purse, B.V., Mellor, P.S. & Rogers, D.J. (2005) Climate change and the recent emergence of bluetongue in Europe. *Nature Reviews Microbiology*, **3**, 171–181.

Ramsdale, C. & Snow, K. (1995) *Mosquito Control in Britain*. University of East London.

Ramsdale, C.D. & Gunn, N. (2005) History of and prospects for mosquito-borne disease in Britain. *European Mosquito Bulletin*, **20**, 15–30.

Randolph, S.E. (2010) Human activities predominate in determining changing incidence of tick-borne encephalitis in Europe. *Euro Surveillance*, **15**, 24–31.

Randolph, S.E. & Rogers, D.J. (2010) The arrival, establishment and spread of exotic diseases: patterns and predictions. *Nature Reviews Microbiology*, **8**, 361–71.

Raxworthy, C.J., Martinez-Meyer, E., Horning, N., Nussbaum, R.A., Schneider, G.E., Ortega-Huerta, M.A. & Townsend Peterson, A. (2003) Predicting distributions of known and unknown reptile species in Madagascar. *Nature*, **426**, 837–41.

Reisen, W.K. (2010) Landscape epidemiology of vector-borne diseases. *Annual Review of Entomology*, **55**, 461–83.

Reiter, P. (2010a) West Nile virus in Europe: understanding the present to gauge the future. *Euro Surveillance*, pp. 1–7.

Reiter, P. & Sprenger, D. (1987) The used tire trade: a mechanism for the worldwide dispersal of container breeding mosquitoes. *Journal of the American Mosquito Control Association*, **3**, 494–501.

Reiter, P. (2010b) Yellow fever and dengue: a threat to Europe? *Eurosurveillance*, **15**, 19509.

Rogers, D.J. (2006) Models for vectors and vector-borne diseases. *Advances in Parasitology*, **62**, 1–35.

Stephens, C.R., Heau, J.G., González, C., Ibarra-Cerdeña, C.N., Sánchez-Cordero, V. & González-Salazar, C. (2009) Using biotic interaction networks for prediction in biodiversity and emerging diseases. *PLoS ONE*, **4**, e5725.

Svobodova, V. & Misonova, P. (2005) The potential risk of Dirofilaria immitis becoming established in the Czech Republic by imported dogs. *Veterinary parasitology*, **128**, 137–40.

Tatem, A.J., Hay, S.I. & Rogers, D.J. (2006) Global traffic and disease vector dispersal. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 6242–7.

Ward, G. (2007) *Statistics in ecological modeling; Presence-only data and Boosted MARS*. Ph.D. thesis, Stanford University.

Ward, G., Hastie, T., Barry, S., Elith, J. & Leathwick, J.R. (2009) Presence-only data and the em algorithm. *Biometrics*, **65**, 554–63.

Warren, D.L. (2012) In defense of 'niche modeling'. *Trends in Ecology & Evolution*, **27**, 497–500.

Wilson, A.J. & Mellor, P.S. (2009) Bluetongue in Europe: past, present and future. *Philosophical Transactions of the Royal Society of London Series B, Biological sciences*, **364**, 2669–81.

Wisz, M.S., Pottier, J., Kissling, W.D., Pellissier, L., Lenoir, J., Damgaard, C.F., Dormann, C.F., Forchhammer, M.C., Grytnes, J.A., Guisan, A., Heikkinen, R.K., Hø ye, T.T., Kühn, I., Luoto, M., Maiorano, L., Nilsson, M.C., Normand, S., Ockinger, E., Schmidt, N.M., Termansen, M., Timmermann, A., Wardle, D.A., Aastrup, P. & Svenning, J.C. (2013) The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biological Reviews of the Cambridge Philosophical Society*, **88**, 15–30.

Chapter 2

West Nile virus vector *Culex modestus* established in southern England

Nick Golding, Miles A. Nunn, Jolyon M. Medlock,
Bethan V. Purse, Alex G. C. Vaux & Stefanie M. Schäfer

Authors' contributions: NG designed and carried out the larval surveys, identified the larvae and wrote the manuscript. SMS initially identified the larvae as *Cx. modestus* and carried out the phylogenetic analysis. JMM and AGCV designed and carried out the adult trapping and identified the adults. MAN and BVP contributed to the design of the larval survey and interpretation of the results. All co-authors contributed to revision of the manuscript.

Abstract

Background: The risk posed to the United Kingdom by West Nile virus (WNV) has previously been considered low, due to the absence or scarcity of the main *Culex* sp. bridge vectors. The mosquito *Culex modestus* is widespread in southern Europe, where it acts as the principle bridge vector of WNV. This species was not previously thought to be present in the United Kingdom.

Findings: Mosquito larval surveys carried out in 2010 identified substantial populations of *Cx. modestus* at two sites in marshland in southeast England. Host-seeking-adult traps placed at a third site indicate that the relative seasonal abundance of *Cx. modestus* peaks in early August. DNA barcoding of these specimens from the United Kingdom and material from southern France confirmed the morphological identification.

Conclusions: *Cx. modestus* appears to be established in the North Kent Marshes, possibly as the result of a recent introduction. The addition of this species to the United Kingdom's mosquito fauna may increase the risk posed to the United Kingdom by WNV.

Findings

Culex modestus is a competent laboratory vector of West Nile virus (WNV, Balenghien *et al.* (2008)) and regularly bites birds, humans and horses in continental Europe (Balenghien, 2006). This mosquito is considered the principle bridge vector of WNV between birds and humans in the Camargue wetland, southern France and is thought to have played a role in the transmission of WNV in the Danube delta, Caspian and Asov sea deltas, and the Volga region in Russia (Ponçon *et al.*, 2007). It has also been implicated in Tahyna and Lednice virus transmission in France and Slovakia respectively (Lundström, 1994).

Cx. modestus is widely distributed in the Palaearctic region, the larvae inhabit fresh to slightly saline water in irrigation channels, marshes and rice fields (Becker *et al.*, 2010). Prior to this report, the only record of this species in the United Kingdom totalled three adults and ten larvae found in and around Portsmouth in southern England in 1944-45 (Marshall, 1945).

The Study

Mosquito surveys were carried out during 2010 in the North Kent Marshes, south-east England (Fig. 3.1). Larval surveys were undertaken at two sites - Cliffe marshes (Cliffe; 51°28'58"N 0°28'45"E) and Elmley National Nature Reserve (Elmley; 51°23'03"N 0°47'19"E) - in June, July and August. At each visit larvae were sampled twice using a 1 litre dipper at randomly located points along the edges of drainage ditches, reed beds and pools. A total of 230 points were sampled, across an area of 3.83km^2. The relative seasonal abundance of host-seeking female mosquitoes was measured at Northward Hill bird reserve (51°27'45"N 0°33'02"E) using a Mosquito Magnet trap (Liberty plus model, American Biophysics, Rhode Island, USA). This site is 5km from Cliffe and 18km from Elmley (Fig. 3.1A). The trap ran for four nights on alternate weeks between April and October.

Larval and adult mosquitoes were identified morphologically using a range of keys (Becker *et al.*, 2010; Snow, 1990; Cranston *et al.*, 1987; Lechthaler, 2005). To confirm the morphological identification DNA barcodes of a subset of *Cx. modestus* specimens from the North Kent Marshes (7 larvae, 10 adults) and the Camargue (3 adults) were generated. A 709 bp fragment of the cytochrome c oxidase subunit I (COI) was amplified by PCR (Folmer *et al.*, 1994) and sequenced. *Cx. modestus* COI barcodes were compared to those obtained from *Cx. pipiens* specimens from the North Kent Marshes ($n = 3$) and Somerset ($n = 6$) as well as 35 COI sequences downloaded from GenBank. Phylogenetic analyses were carried out using MEGA5 software (Tamura *et al.*, 2011).

In larval surveys 850 *Cx. modestus* of all stages were collected, along with *Anopheles maculipennis* s.l., *Cx. pipiens* s.l. and *Culiseta annulata*. At both sites *Cx. modestus* was the second most abundant species after *Cx. pipiens* s.l., making up 44% and 23% of the overall larval population sampled at Cliffe and Elmley respectively (Table 2.1).

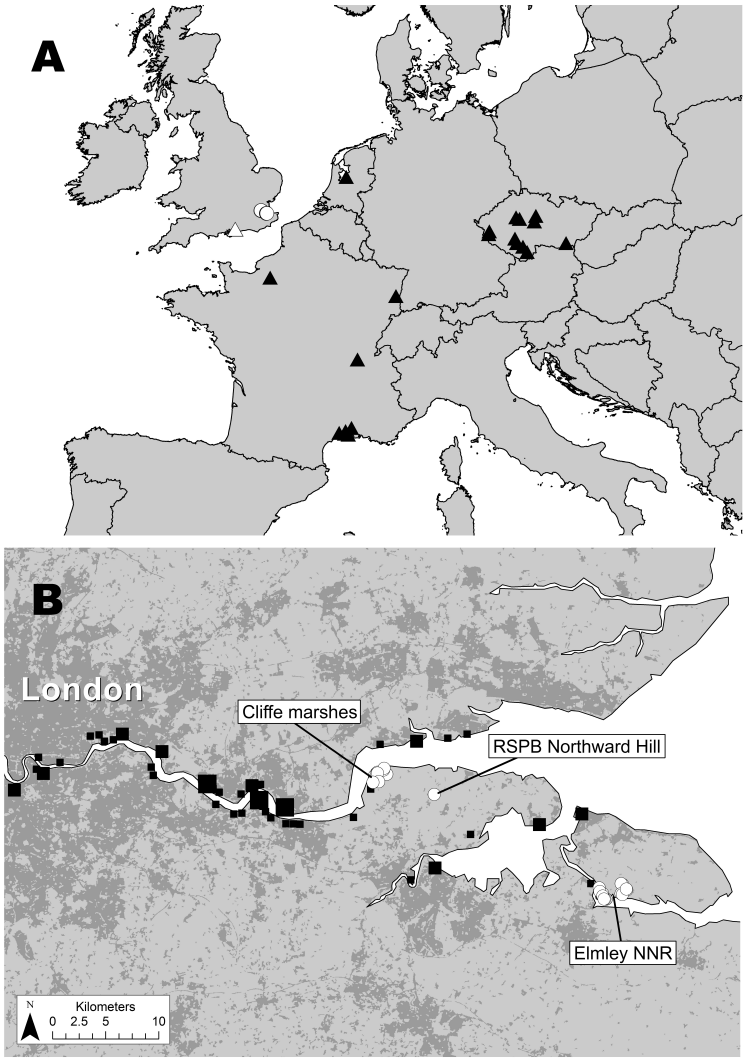


Fig. 2.1: **A)** North-west Europe, showing locations where *Culex modestus* populations were detected in this study (white circles), the location of *Cx. modestus* identified in southern England in 1944-5 (white triangle), and recent records from Europe (black triangles; Francis Schaffner, personal communication and articles cited here). All of these recent records date from the period 2004-2009 with the exception of the two northernmost French records, which date from 1995 and 1998. **B)** Thames Estuary area, showing locations where *Cx. modestus* populations were detected in 2010 (white circles) and the locations of international shipping terminals (black squares). To give an indication of the size of the port, black squares are proportional to number of ships arriving during May 2011: small squares 1-25 ships; medium squares 26-75 ships; large squares 75-165 ships. Urban and semi-urban areas, as classified by the United Kingdom Land Cover Map 2000, are coloured dark grey.

Table 2.1: Numbers and proportions (given as %) of larvae collected from Cliffe marshes and Elmley National Nature Reserve

	Month	*Culex modestus*	%	*Culex pipiens s.l.*	%	*Culiseta annulata*	%	*Anopheles maculipennis s.l.*	%
Cliffe marshes	June	0	-	0	-	0	-	0	-
	July	131	61	52	24	0	-	31	15
	August	231	38	351	58	0	-	23	4
	Total	**362**	**44**	**403**	**49**	**0**	**-**	**54**	**7**
Elmley NNR	June	2	3	57	90	1	2	3	5
	July	371	44	408	49	22	3	36	4
	August	120	10	637	52	377	31	91	7
	Total	**493**	**23**	**1102**	**52**	**400**	**19**	**130**	**6**

Table 2.2: Numbers and proportions (given as %) of adult female *Cx. modestus* collected at RSPB Northward Hill

	14-18 June	12-16 July	26-30 July	09-13 Aug	23-27 Aug	06-10 Sept	20-24 Sept
Culex modestus	0	31	272	325	11	10	0
Total catch	0	120	281	350	40	49	21
%	-	26	97	93	28	20	0

A total of 649 adult female *Cx. modestus* were captured at Northward Hill between 12 July and 10 September, with a peak of 325 adults in the second week of August (Table 2.2). Overall, *Cx. modestus* comprised 75% of the mosquitoes collected at Northward Hill. Morphological identification of *Cx. modestus* was confirmed by DNA barcoding and phylogenetic analyses. All the COI sequences from *Cx. modestus* specimens form a discrete clade with high bootstrap support (Fig. 2.2).

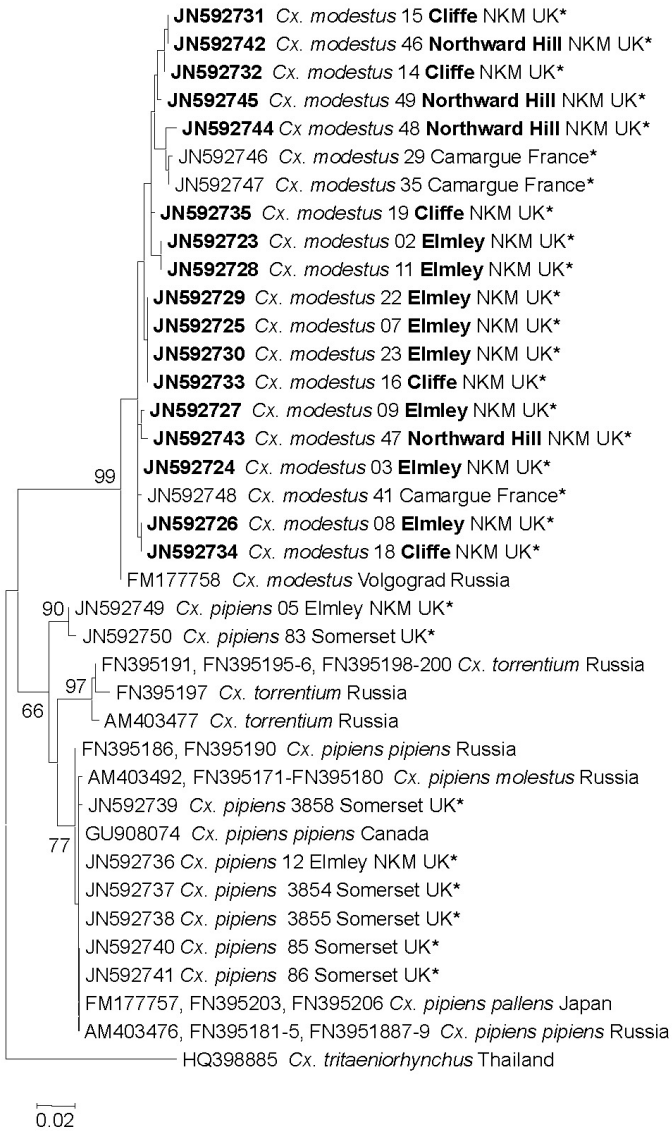


Fig. 2.2: Phylogenetic maximum-likelihood tree of COI sequences (603 bp) from *Culex modestus* (North Kent Marshes (NKM) specimens in boldface) and other representatives of the *Culex* genus estimated using the T92+Γ+I model of nucleotide substitution, which was selected by MODELTEST. Bootstrap values are shown for the main clades only. The accession numbers and geographic origin of the 35 GenBank downloads (which represent 10 unique sequences) and COI barcodes generated in this study (asterisked) are shown. Scale bar indicates nucleotide substitutions per site.

Conclusions

Established populations of *Cx. modestus* have been reported from the Camargue and Dombes wetlands in southern and central France (Ponçon *et al.*, 2007; Pradel *et al.*, 2009) as well as in wetlands in the Czech Republic (Votýpka *et al.*, 2008) but the species is believed to be more widely distributed than this in Europe. Its previous known northerly limit in Europe was in northern France (see Fig. 3.1A, Francis Schaffner, personal communication) and Oostvardersplassen, the Netherlands (Reusken *et al.*, 2010). However these records comprise only a few specimens and it is unclear whether there are established populations at these sites. The species was not detected during a recent and intensive survey of the mosquito fauna of Belgium (Van Bortel *et al.*, 2009). Our finding demonstrates that established populations of *Cx. modestus* are present in the United Kingdom and provides further support for the existence of northern populations of the species.

It seems unlikely that *Cx. modestus* could have been present in the North Kent Marshes for a long time without being detected. The mosquito fauna of the North Kent Marshes are among the most well sampled in the United Kingdom, both by amateur entomologists and by professionals engaged in mosquito control (Ramsdale & Snow, 1995).

An extensive larval survey was carried out at Elmley in 2003 (Hutchinson, 2004). This survey identified 95 sites containing *An. maculipennis* s.l. but did not detect *Cx. modestus*. In the present larval survey *Cx. modestus* were found to be strongly associated with *An. maculipennis* s.l.; being present in 73% of sites containing *An. maculipennis* s.l. larvae. This suggests *Cx. modestus* was absent from this site in 2003. However the 2003 survey did not record any *Cx. pipiens* s.l. in *An. maculipennis* s.l. positive sites, whilst they were present in 20% of such sites in the present study; suggesting that the sampling strategy employed in 2003 may not have been sensitive to culicines.

If these *Cx. modestus* populations were established recently, international shipping

may well have been the route of introduction. International shipping has previously been implicated in the introduction of mosquito species; including *Cx. modestus* to China (Nie *et al.*, 2004) and there are a high number of shipping terminals in the area of the North Kent Marshes (see Fig. 3.1B).

A number of vectors and vector-borne diseases have undergone changes in their geographic ranges in recent years, in response to varied biotic and abiotic environmental factors (Randolph & Rogers, 2010). There is some evidence that *Cx. modestus* is extending its distribution in Europe, with speculation that this may be driven by weather events or changes to wetlands (Pradel *et al.*, 2009; Votýpka *et al.*, 2008). Without detailed information on the previous and current distribution of this species, however, it is unclear what role these factors might play.

A recent review of the potential vectors of WNV (Medlock *et al.*, 2005) concluded that the risk of human cases in the United Kingdom is low due to limited human exposure to potential bridge vectors. However, the risk of transmission of WNV in this part of Kent may be higher than previously supposed, as we have shown that *Cx. modestus* populations exist alongside the potential WNV maintenance vector *Cx. pipiens* s.l. at sites hosting many migratory and resident birds. Since human population numbers in the North Kent Marshes are relatively low and little is known of the dispersal range or host preferences of *Cx. modestus* in the United Kingdom, it is difficult to quantify the significance of any change in risk to humans. It does seem likely, however, that the risk posed to horses, which are often grazed in the North Kent Marshes, will have increased. In light of this, and until the national distribution of *Cx. modestus* is established, surveillance for WNV in the United Kingdom should now focus on this part of Kent.

In summary, the discovery of populations of *Cx. modestus* in southern England suggests a recent introduction of this species and provides further evidence for expansion of its geographic range. There is an associated increased risk posed to the United Kingdom by WNV and other pathogens transmitted by *Cx. modestus*.

Acknowledgements

We thank Steve Gordon and Andy Daw for their cooperation and help with fieldwork, Stephen Larcombe for providing *Cx. modestus* from the Camargue, and the Port of London Authority for providing shipping data. NG, SS, MN and BVP acknowledge funding from the NERC Centre for Ecology & Hydrology (CEH Environmental Change Integrating Fund programme). JM and AV are funded by HPA Government Grant-in-Aid and an HPA fund for nationwide mosquito surveillance.

References

Balenghien, T. (2006) *De l'identification des vecteurs du virus West Nile à la modélisation du risque d'infection dans le sud de la France.* Ph.D. thesis, lUniversité des Sciences des Technologies et de la Santé de Grenoble.

Balenghien, T., Vazeille, M., Grandadam, M., Schaffner, F., Zeller, H., Reiter, P., Sabatier, P., Fouque, F. & Bicout, D.J. (2008) Vector competence of some French Culex and Aedes mosquitoes for West Nile virus. *Vector-Borne and Zoonotic Diseases*, **8**, 589–95.

Becker, N., Petric, D., Zgomba, M., Boase, C., Madon, M., Dahl, C. & Kaiser, A. (2010) *Mosquitoes and Their Control.* Springer Verlag, Berlin, second edition.

Cranston, P., Ramsdale, C.D., Snow, K.R. & White, G. (1987) *Adults, Larvae, and Pupae of British Mosquitoes (Culicidae) A Key.* Freshwater Biological Association.

Folmer, O., Black, M., Hoeh, W., Lutz, R. & Vrijenhoek, R. (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, **3**, 294–9.

Hutchinson, R.A. (2004) *Mosquito Borne Diseases in England: past, present and future risks, with special reference to malaria in the Kent Marshes.* Ph.D. thesis, Durham.

Lechthaler, W. (2005) *Culicidae 05 - Key to Larvae, Pupae and Males from Central and Western Europe. (CD-edition).*

Lundström, J.O. (1994) Vector competence of Western European mosquitoes for arboviruses: A review of field and experimental studies. *Bulletin of the Society of Vector Ecologists*, **19**, 23–36.

Marshall, J.F. (1945) Records of Culex (Barraudius) modestus Ficalbi (Diptera, Culicidæ) obtained in the South of England. *Nature*, **156**, 172–173.

Medlock, J.M., Leach, S. & Snow, K.R. (2005) Potential transmission of West Nile virus in the British Isles: an ecological review of candidate mosquito bridge vectors. *Medical and Veterinary Entomology*, **19**, 2–21.

Nie, W.Z., Li, J.C., Li, D.X., Wang, R.J. & Gratz, N.G. (2004) Mosquitoes found aboard ships arriving at Qinhuangdao Port, P. R. China. *Medical Entomology and Zoology*, **55**, 333–335.

Ponçon, N., Balenghien, T., Toty, C., Baptiste Ferré, J., Thomas, C., Dervieux, A., L'Ambert, G., Schaffner, F., Bardin, O. & Fontenille, D. (2007) Effects of Local Anthropogenic Changes on Potential Malaria Vector Anopheles hyrcanus and West Nile Virus Vector Culex modestus, Camargue, France. *Emerging Infectious Diseases*, **13**, 1810–5.

Pradel, J.A., Martin, T., Rey, D., Foussadier, R. & Bicout, D.J. (2009) Is Culex modestus (Diptera: Culicidae), Vector of West Nile Virus, Spreading in the Dombes Area, France? *Journal of Medical Entomology*, **46**, 1269–1281.

Ramsdale, C. & Snow, K. (1995) *Mosquito Control in Britain.* University of East London.

Randolph, S.E. & Rogers, D.J. (2010) The arrival, establishment and spread of exotic diseases: patterns and predictions. *Nature Reviews Microbiology*, **8**, 361–71.

Reusken, C., De Vries, A., Den Hartog, W., Braks, M. & Scholte, E.J. (2010) A study of the circulation of West Nile virus in mosquitoes in a potential high-risk area for arbovirus circulation in the Netherlands,De Oostvaardersplassen. *European Mosquito Bulletin*, **28**, 69–83.

Snow, K.R. (1990) *Mosquitoes (Naturalists' Handbooks 14).* Richmond Publishing Company.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evo-

lutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, **28**, 2731–2739.

Van Bortel, W., Grootaert, P., Hance, T., Hendrickx, G. & Takken, W. (2009) Mosquito Vectors of Disease: Spatial Biodiversity, Drivers of Change and Risk "MODIRISK" Final Report Phase 1. Technical report, Brussels.

Votýpka, J., Šeblová, V. & Rádrová, J. (2008) Spread of the West Nile virus vector Culex modestus and the potential malaria vector Anopheles hyrcanus in central Europe. *Journal of Vector Ecology*, **33**, 269–277.

Appendix 2.A Print version

Golding et al. Parasites & Vectors 2012, 5:32
http://www.parasitesandvectors.com/content/5/1/32

Parasites & Vectors

SHORT REPORT Open Access

West Nile virus vector *Culex modestus* established in southern England

Nick Golding[1,2*], Miles A Nunn[2], Jolyon M Medlock[3], Bethan V Purse[4], Alexander GC Vaux[3] and Stefanie M Schäfer[2]

Abstract

Background: The risk posed to the United Kingdom by West Nile virus (WNV) has previously been considered low, due to the absence or scarcity of the main *Culex* sp. bridge vectors. The mosquito *Culex modestus* is widespread in southern Europe, where it acts as the principle bridge vector of WNV. This species was not previously thought to be present in the United Kingdom.

Findings: Mosquito larval surveys carried out in 2010 identified substantial populations of *Cx. modestus* at two sites in marshland in southeast England. Host-seeking-adult traps placed at a third site indicate that the relative seasonal abundance of *Cx. modestus* peaks in early August. DNA barcoding of these specimens from the United Kingdom and material from southern France confirmed the morphological identification.

Conclusions: *Cx. modestus* appears to be established in the North Kent Marshes, possibly as the result of a recent introduction. The addition of this species to the United Kingdom's mosquito fauna may increase the risk posed to the United Kingdom by WNV.

Keywords: Anopheles, Arboviruses, Culex, Culicidae, Disease Vectors, DNA Barcoding, Taxonomic, Introduced Species, West Nile virus

Findings

Culex modestus is a competent laboratory vector of West Nile virus (WNV, [1]) and regularly bites birds, humans and horses in continental Europe [2]. This mosquito is considered the principle bridge vector of WNV between birds and humans in the Camargue wetland, southern France and is thought to have played a role in the transmission of WNV in the Danube delta, Caspian and Asov sea deltas, and the Volga region in Russia [3]. It has also been implicated in Tahyna and Lednice virus transmission in France and Slovakia respectively [4].

Cx. modestus is widely distributed in the Palaearctic region, the larvae inhabit fresh to slightly saline water in irrigation channels, marshes and rice fields [5]. Prior to this report, the only record of this species in the United Kingdom totalled three adults and ten larvae found in and around Portsmouth in southern England in 1944-45 [6].

The Study

Mosquito surveys were carried out during 2010 in the North Kent Marshes, south-east England (Figure 1). Larval surveys were undertaken at two sites - Cliffe marshes (Cliffe; 51°28'58"N 0°28'45"E) and Elmley National Nature Reserve (Elmley; 51°23'03"N 0°47'19"E) - in June, July and August. At each visit larvae were sampled twice using a 1 litre dipper at randomly located points along the edges of drainage ditches, reed beds and pools. A total of 230 points were sampled, across an area of 3.83 km^2. The relative seasonal abundance of host-seeking female mosquitoes was measured at Northward Hill bird reserve (51°27'45"N 0°33'2"E) using a Mosquito Magnet trap (Liberty plus model, American Biophysics, Rhode Island, USA). This site is 5 km from Cliffe and 18 km from Elmley (Figure 1A). The trap ran for four nights on alternate weeks between April and October.

* Correspondence: nick.golding@zoo.ox.ac.uk
[1]Spatial Ecology and Epidemiology Group, Tinbergen Building, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK
Full list of author information is available at the end of the article

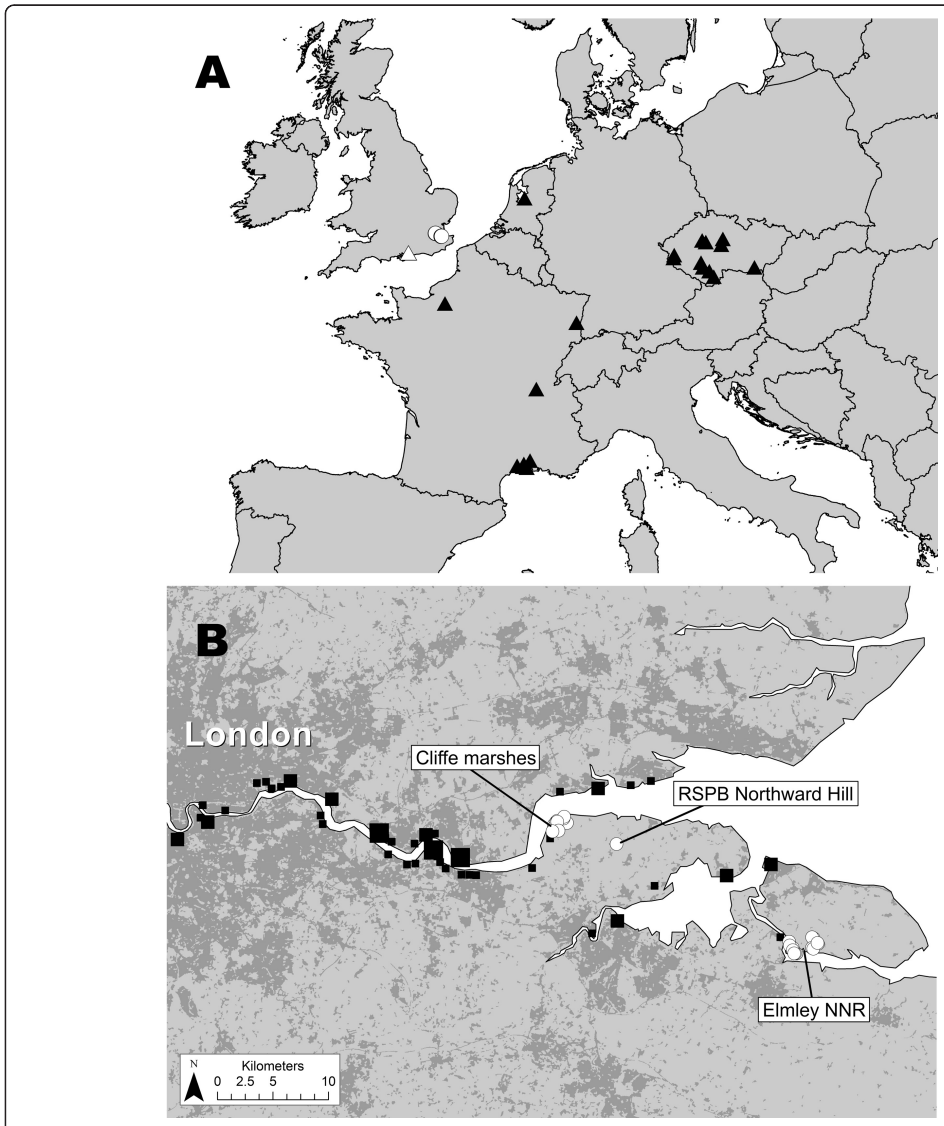


Figure 1 A) North-west Europe, showing locations where *Culex modestus* populations were detected in this study (white circles), the location of *Cx. modestus* identified in southern England in 1944-5 (white triangle), and recent records from Europe (black triangles; Francis Schaffner, personal communication and articles cited here). All of these recent records date from the period 2004-2009 with the exception of the two northernmost French records, which date from 1995 and 1998. B) Thames Estuary area, showing locations where *Cx. modestus* populations were detected in 2010 (white circles) and the locations of international shipping terminals (black squares). To give an indication of the size of the port, black squares are proportional to number of ships arriving during May 2011: small squares 1-25 ships; medium squares 26-75 ships; large squares 75-165 ships. Urban and semi-urban areas, as classified by the United Kingdom Land Cover Map 2000, are coloured dark grey.

Larval and adult mosquitoes were identified morphologically using a range of keys [5,7-9]. To confirm the morphological identification DNA barcodes of a subset of *Cx. modestus* specimens from the North Kent Marshes (7 larvae, 10 adults) and the Camargue (3 adults) were generated. A 709 bp fragment of the cytochrome c oxidase subunit I (COI) was amplified by PCR [10] and sequenced. *Cx. modestus* COI barcodes were

compared to those obtained from *Cx. pipiens* specimens from the North Kent Marshes (n = 3) and Somerset (n = 6) as well as 35 COI sequences downloaded from GenBank. Phylogenetic analyses were carried out using MEGA5 software [11].

In larval surveys 850 *Cx. modestus* of all stages were collected, along with *Anopheles maculipennis* s.l., *Cx. pipiens* s.l. and *Culiseta annulata*. At both sites *Cx. modestus* was the second most abundant species after *Cx. pipiens* s.l., making up 44% and 23% of the overall larval population sampled at Cliffe and Elmley respectively (Table 1).

A total of 649 adult female *Cx. modestus* were captured at Northward Hill between 12 July and 10 September, with a peak of 325 adults in the second week of August (Table 2). Overall, *Cx. modestus* comprised 75% of the mosquitoes collected at Northward Hill. Morphological identification of *Cx. modestus* was confirmed by DNA barcoding and phylogenetic analyses. All the COI sequences from *Cx. modestus* specimens form a discrete clade with high bootstrap support (Figure 2).

Conclusions

Established populations of *Cx. modestus* have been reported from the Camargue and Dombes wetlands in southern and central France [3,12] as well as in wetlands in the Czech Republic [13] but the species is believed to be more widely distributed than this in Europe. Its previous known northerly limit in Europe was in northern France (see Figure 1A, Francis Schaffner, personal communication) and Oostvardersplassen, the Netherlands [14]. However these records comprise only a few specimens and it is unclear whether there are established populations at these sites. The species was not detected during a recent and intensive survey of the mosquito fauna of Belgium [15]. Our finding demonstrates that established populations of *Cx. modestus* are present in the United Kingdom and provides further support for the existence of northern populations of the species.

It seems unlikely that *Cx. modestus* could have been present in the North Kent Marshes for a long time without being detected. The mosquito fauna of the North Kent Marshes are among the most well sampled in the United Kingdom, both by amateur entomologists and by professionals engaged in mosquito control [16].

An extensive larval survey was carried out at Elmley in 2003 [17]. This survey identified 95 sites containing *An. maculipennis* s.l. but did not detect *Cx. modestus*. In the present larval survey *Cx. modestus* were found to be strongly associated with *An. maculipennis* s.l.; being present in 73% of sites containing *An. maculipennis* s.l. larvae. This suggests *Cx. modestus* was absent from this site in 2003. However the 2003 survey did not record any *Cx. pipiens* s.l. in *An. maculipennis* s.l. positive sites, whilst they were present in 20% of such sites in the present study; suggesting that the sampling strategy employed in 2003 may not have been sensitive to Culicines.

If these *Cx. modestus* populations were established recently, international shipping may well have been the route of introduction. International shipping has previously been implicated in the introduction of mosquito species; including *Cx. modestus* to China [18] and there are a high number of shipping terminals in the area of the North Kent Marshes (see Figure 1B).

A number of vectors and vector-borne diseases have undergone changes in their geographic ranges in recent years, in response to varied biotic and abiotic environmental factors [19]. There is some evidence that *Cx. modestus* is extending its distribution in Europe, with speculation that this may be driven by weather events or changes to wetlands [12,13]. Without detailed information on the previous and current distribution of this species, however, it is unclear what role these factors might play.

A recent review of the potential vectors of WNV [20] concluded that the risk of human cases in the United Kingdom is low due to limited human exposure to potential bridge vectors. However, the risk of transmission of WNV in this part of Kent may be higher than

Table 1 Numbers and proportions (given as %) of larvae collected from Cliffe marshes and Elmley National Nature Reserve

	Month	*Culex modestus*	%	*Culex pipiens s.l.*	%	*Culiseta annulata*	%	*Anopheles maculipennis s.l.*	%
Cliffe marshes	June	0		0		0		0	
	July	131	61	52	24	0	0	31	15
	August	231	38	351	58	0	0	23	4
	Total	**362**	**44**	**403**	**49**	**0**	**0**	**54**	**7**
Elmley NNR	June	2	3	57	90	1	2	3	5
	July	371	44	408	49	22	3	36	4
	August	120	10	637	52	377	31	91	7
	Total	**493**	**23**	**1102**	**52**	**400**	**19**	**130**	**6**

Table 2 Numbers and proportions (given as %) of adult female *Cx. modestus* collected at RSPB Northward Hill

	14-18 June	12-16 July	26-30 July	09-13 August	23-27 August	06-10 September	20-24 September
Cx. modestus	0	31	272	325	11	10	0
Total catch	0	120	281	350	40	49	21
%	-	26	97	93	28	20	0

previously supposed, as we have shown that *Cx. modestus* populations exist alongside the potential WNV maintenance vector *Cx. pipiens* s.l. at sites hosting many migratory and resident birds. Since human population numbers in the North Kent Marshes are relatively low and little is known of the dispersal range or host preferences of *Cx. modestus* in the United Kingdom, it is difficult to quantify the significance of any change in risk to humans. It does seem likely, however, that the risk posed to horses, which are often grazed in the North Kent Marshes, will have increased. In light of this, and until the national distribution of *Cx. modestus* is established, surveillance for WNV in the United Kingdom should now focus on this part of Kent.

In summary, the discovery of populations of *Cx. modestus* in southern England suggests a recent introduction of this species and provides further evidence for expansion of its geographic range. There is an associated increased risk posed to the United Kingdom by WNV and other pathogens transmitted by *Cx. modestus*.

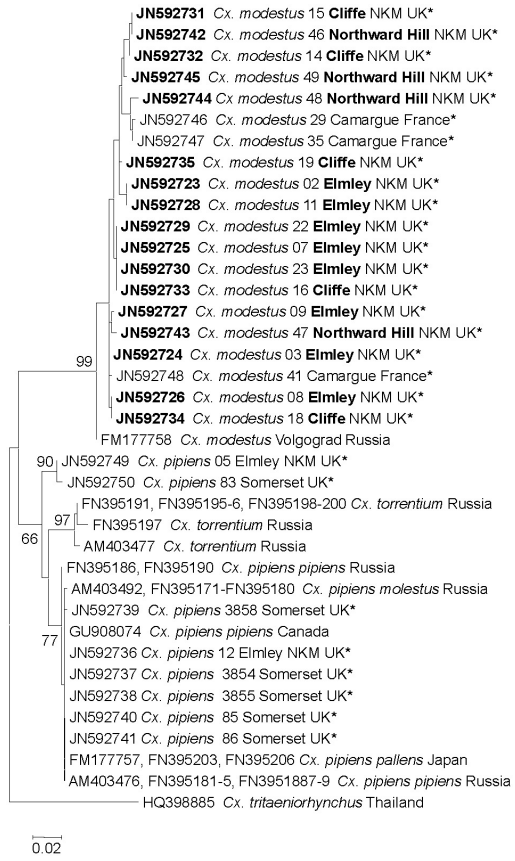


Figure 2 Phylogenetic maximum-likelihood tree of COI sequences (603 bp) from *Culex modestus* (North Kent Marshes (NKM) specimens in boldface) and other representatives of the *Culex* genus estimated using the T92+Γ+I model of nucleotide substitution, which was selected by MODELTEST. Bootstrap values are shown for the main clades only. The accession numbers and geographic origin of the 35 GenBank downloads (which represent 10 unique sequences) and COI barcodes generated in this study (asterisked) are shown. Scale bar indicates nucleotide substitutions per site.

Acknowledgements
We thank Steve Gordon and Andy Daw for their cooperation and help with fieldwork, Stephen Larcombe for providing *Cx. modestus* from the Camargue, and the Port of London Authority for providing shipping data. NG, SS, MN, and BVP acknowledge funding from the NERC Centre for Ecology & Hydrology (CEH Environmental Change Integrating Fund programme). JM and AV are funded by HPA Government Grant-in-Aid, and an HPA fund for nationwide mosquito surveillance.

Author details
[1]Spatial Ecology and Epidemiology Group, Tinbergen Building, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK. [2]Centre for Ecology & Hydrology, Maclean Building, Benson Lane, Crowmarsh Gifford, Wallingford OX10 8BB, UK. [3]Medical Entomology & Zoonoses Ecology group, Microbial Risk Assessment, Emergency Response Department, Health Protection Agency, Porton Down, Salisbury, Wiltshire SP4 0JG, UK. [4]Centre for Ecology & Hydrology, Bush Estate, Penicuik, Midlothian, EH26 0QB, UK.

Authors' contributions
NG designed and carried out the larval surveys, identified the larvae and drafted the manuscript. SMS initially identified the larvae as *Cx. modestus*, carried out the phylogenetic analysis and contributed to revision of the manuscript. JMM and AGCV designed and carried out the adult trapping, identified the adults and contributed to revision of the manuscript. MAN and BVP contributed to the design of the larval survey, interpretation of the results and revision of the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 10 November 2011 Accepted: 9 February 2012
Published: 9 February 2012

References

1. Balenghien T, Vazeille M, Grandadam M, Schaffner F, Zeller H, Reiter P, Sabatier P, Fouque F, Bicout DJ: **Vector competence of some French**

Culex and Aedes mosquitoes for West Nile virus. *Vector-Borne Zoonot Dis* 2008, **8**:589-95.
2. Balenghien T, Fouque F, Sabatier P, Bicout DJ: **Horse-, Bird-, and Human-Seeking Behaviour and Seasonal Abundance of Mosquitoes in a West Nile Virus Focus of Southern France.** *J Med Entomol* 2006, **43**:936-946.
3. Ponçon N, Balenghien T, Toty C, Baptiste Ferré J, Thomas C, Dervieux A, L'ambert G, Schaffner F, Bardin O, Fontenille D: **Effects of Local Anthropogenic Changes on Potential Malaria Vector Anopheles hyrcanus and West Nile Virus Vector Culex modestus, Camargue, France.** *Emerg Infect Dis* 2007, **13**:1810-5.
4. Lundström JO: **Vector competence of Western European mosquitoes for arboviruses: A review of field and experimental studies.** *Bull Soc Vector Ecol* 1994, **19**:23-36.
5. Becker N, Petric D, Zgomba M, Boase C, Madon M, Dahl C, Kaiser A: *Mosquitoes and their control* Second. Berlin: Springer Verlag; 2010.
6. Marshall JF: **Records of Culex (Barraudius) modestus Ficalbi (Diptera, Culicidæ) obtained in the South of England.** *Nature* 1945, **156**:172-173.
7. Snow KR: *Mosquitoes (Naturalists' Handbooks 14)* Richmond Publishing; 1990.
8. Cranston PS, Ramsdale CD, Snow KR, White GB: *Adults, Larvae, and Pupae of British Mosquitoes (Culicidae) A Key* Freshwater Biological Association; 1987.
9. Lechthaler W: *Culicidae 05 - Key to Larvae, Pupae and Males from Central and Western Europe. (CD-edition). Eutaxa* 2005.
10. Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R: **DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates.** *Mol Mar Biol Biotech* 1994, **3**:294-9.
11. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.** *Mol Biol Evol* 2011, **28**:2731-2739.
12. Pradel JA, Martin T, Rey D, Foussadier R, Bicout DJ: **Is Culex modestus (Diptera: Culicidae), Vector of West Nile Virus, Spreading in the Dombes Area, France?** *J Med Entomol* 2009, **46**:1269-1281.
13. Votýpka J, Šeblová V, Rádrová J: **Spread of the West Nile virus vector Culex modestus and the potential malaria vector Anopheles hyrcanus in central Europe.** *J Vector Ecol* 2008, **33**:269-277.
14. Reusken C, De Vries A, Den Hartog W, Braks M, Scholte E-J: **A study of the circulation of West Nile virus in mosquitoes in a potential high-risk area for arbovirus circulation in the Netherlands,"De Oostvaardersplassen".** *Europ Mosq Bull* 2010, **28**:69-83.
15. Van Bortel W, Grootaert P, Hance T, Hendrickx G, Takken W: *Mosquito Vectors of Disease: Spatial Biodiversity, Drivers of Change and Risk "MODIRISK" Final Report Phase 1* Brussels: Belgian Science Policy; 2009.
16. Ramsdale C, Snow KR: *Mosquito control in Britain* London: University of East London Press; 1995.
17. Hutchinson RA: **Mosquito Borne Diseases in England: past, present and future risks, with special reference to malaria in the Kent Marshes.** *PhD Thesis* University of Durham, Department of Biological & Biomedical Sciences; 2004.
18. Nie W-Z, Li J-C, Li D-X, Wang R-J, Gratz N: **Mosquitoes found aboard ships arriving at Qinhuangdao Port, P. R. China.** *Med Entomol Zool* 2004, **55**:333-335.
19. Randolph SE, Rogers DJ: **The arrival, establishment and spread of exotic diseases: patterns and predictions.** *Nat Rev Microbiol* 2010, **8**:361-71.
20. Medlock J, Leach S, Snow KR: **Potential transmission of West Nile virus in the British Isles: an ecological review of candidate mosquito bridge vectors.** *Med Vet Entomol* 2005, **19**:2-21.

doi:10.1186/1756-3305-5-32
Cite this article as: Golding *et al.*: West Nile virus vector *Culex modestus* established in southern England. *Parasites & Vectors* 2012 **5**:32.

Chapter 3

Identifying environmental conditions and biotic interactions driving the larval community composition of vector mosquitoes using a Bayesian multivariate-binomial distribution model

Nick Golding, Miles A. Nunn & Bethan V. Purse

Authors' contributions: NG designed and carried out the field and laboratory work, devised and implemented the statistical analysis and wrote the manuscript. MAN and BVP contributed to the design of the larval survey, interpretation of the results and revision of the manuscript.

Abstract

Spatial variation in the risk of mosquito-borne pathogens is strongly influenced by the distribution of communities of suitable vector mosquitoes. Individual species' distributions are known to be strongly influenced by their abiotic habitat requirements, but the importance of biotic interactions between species is less well understood. Whilst previous studies have investigated biotic interactions relevant to mosquito communities under laboratory settings, the impact of biotic interactions on mosquito distributions in the field have rarely been assessed.

We construct a novel multivariate-binomial species interaction distribution model (SIDM) to investigate community interactions using observational data from a UK wetland. Mosquito larval distributions in this habitat are strongly driven by environmental covariates, with shallow water positively associated with all mosquito species. Using the model we are able to distinguish between the environmental factors and inter-species interactions influencing the distribution of a community of potential vector mosquitoes. We identify ditch shrimp (genus *Palaemonetes*) and fish as predators which may influence mosquito distributions.

Whilst SIDMs do not yield concrete evidence of biotic interactions, they can suggest potentially important interactions which may benefit from further study. The statistical model we develop has advantages over existing approaches and could be applied widely in community ecology. We provide an R package `BayesComm` to enable its wider use.

Introduction

The spatial distribution of mosquito-borne diseases (MBDs) is heavily dependent on the distribution of suitable vector species (Reisen, 2010). Whilst transmission of some MBDs is entirely dependent on single mosquito species, most may be transmitted by multiple species (Becker *et al.*, 2010). The involvement of multiple vector species with differing ecology and biology can complicate vector control interventions. Understanding and mapping the spatial distribution of these species is therefore essential for efficient control of globally important diseases such as malaria (Hay *et al.*, 2010; Ferguson *et al.*, 2010).

For some zoonotic MBDs, transmission of the pathogen from its sylvatic hosts to human populations requires the presence of multiple species. Such a transmission cycle is exemplified by West Nile virus. WNV is sustained in an avian sylvatic cycle by 'maintenance' mosquito species which must both be ornithophagic and competent for transmission of the virus. Humans and other mammals are 'dead-end' hosts for WNV as they rarely develop a sufficient level of viraemia to re-infect mosquitoes. In order for humans to become infected by the virus they must therefore be bitten by a 'bridge' mosquito species which has previously fed on an infected bird and gone on to establish an infection (and therefore must be ornithophagic, anthropophagic and competent). The risk of human WNV cases is therefore restricted to areas where susceptible avian hosts, human populations, maintenance mosquito species and bridge species coincide in both space and time.

Unsurprisingly, transmission of WNV in Europe exhibits a patchy distribution which appears to principally be driven by the distribution of its major vectors (Durand *et al.*, 2010). In order to identify areas at risk from WNV and similar diseases, both now and in the future, we must therefore understand what drives the assembly and distribution of entire communities of vector mosquitoes

Whilst adult, female mosquitoes are responsible for transmitting pathogens, the ecology of the relatively immobile larval stages drives the distribution of the species

at all but the finest of spatial scales. Herein we refer to the distributions of larval mosquitoes.

Drivers of community assembly

The majority of previous work has focussed on environmental (abiotic) drivers of the distributions of individual mosquito species (Sinka *et al.*, 2010; Diuk-Wasser *et al.*, 2006; Tran *et al.*, 2008) and mosquito communities (Beketov *et al.*, 2010; Steiger *et al.*, 2012; Schäfer *et al.*, 2004; Brown *et al.*, 2008). Less attention has been paid to the impact of between-species (biotic) interactions on vector distributions. There are a number of ways biotic interactions may influence the distribution of mosquitoes. Predation and competition between species may restrict mosquito distributions, whilst apparent or indirect mutualisms may promote them (Blaustein & Chase, 2007). Identifying biotic relationships that affect the distribution of vector mosquitoes could be useful for disease control (Ferguson *et al.*, 2010; Mukabana *et al.*, 2006).

Investigations of biotic interactions affecting mosquitoes have mainly been restricted to laboratory experiments (Juliano, 2009; Blaustein & Chase, 2007; Medlock & Snow, 2008). Whilst laboratory studies allow identification of biotic interactions which are possible under certain environmental conditions, they cannot tell us whether these interactions influence the distributions of mosquito communities in the field. Experimental manipulations of communities in the field would provide a fair test of the impact of specific biotic interactions on the distribution of mosquito communities. Carrying out such experiments at appropriate spatial and temporal scale and with sufficient replicates to draw generalisable conclusions would likely be financially and practically unfeasible, particularly, as here, where there are a large number of potential interactions to test.

By contrast, observational data on species co-occurrences are relatively easy to collect. Such data could be interrogated for signs of biotic interactions between species, manifested as correlations between their distributions. This approach is complicated by the difficulty in distinguishing the effects of biotic interactions from environmental

factors. Positive (or negative) correlations between species' distributions could just as easily be explained by sharing (or not sharing) environmental habitat requirements as by the presence of a biotic interaction. In order to infer biotic interactions from such data, we therefore need appropriate methods to discriminate between biotic and environmental drivers of species distributions

Species interaction distribution models

A number of species interaction distribution models (SIDMs) have been proposed that aim to quantify the effects of biotic interactions on species distributions (Kissling *et al.*, 2011). Of these, multivariate binomial regression provides an appealing conceptual and technical approach (Ovaskainen *et al.*, 2010). In such a model each species' *fundamental niche* is modelled by independent binomial regressions and biotic interactions between species are modelled as a symmetric matrix controlling correlation in the regression errors. Because the model accounts for co-occurrence in distributions which can be explained by each species' fundamental niche, the positive and negative correlation coefficients are assumed to be representative of positive and negative biotic interactions between species. Parameter inference is carried out using Markov Chain Monte Carlo (MCMC) because of the relative complexity of the statistical model.

Here, we investigate the abiotic environmental factors and inter-species interactions influencing the spatial distribution of a community of potential vector mosquitoes in a UK wetland. We identify the most important abiotic environmental drivers of each species' distribution in this habitat. Using a novel and efficient sampler for the multivariate binomial regression model we identify predators of mosquito larvae which appear to influence the distribution of the mosquito vector species. We discuss the implications of these findings and the potential for SIDMs to improve our understanding of the community ecology of vector-borne diseases.

Materials and methods

Larval habitat surveys

Dipping surveys were carried out at three sites in the North Kent Marshes in south-east England; one at Cliffe marshes (51°28'58"N 0°28'45"E) and two at Elmley Marshes (51°23'03"N 0°47'19"E, see Fig. 3.1). A set of random coordinates were determined within each site, the nearest water body to each of these points was marked as a dipping site and its location (accurate to within 10cm) recorded using a differential GPS system (Trimble 5800, Trimble Navigation Limited, Sunnyvale, California, USA). A total of 167 dip sites were each visited six times (rounds); in June, July and August of 2010 and 2011, but few mosquitoes were collected in June of both years, so data from these two rounds were discarded. At every dipping round each of the dip sites were visited and a pair of dips was carried out using a 1 litre dipper; one at the edge and one toward the centre of the water body. The contents of these two dips were pooled for analysis. Mosquito larvae were identified morphologically using the keys of Snow (1990); Cranston *et al.* (1987); Becker *et al.* (2010) and presence or absence for each species at each dip recorded.

Six representative morphologically identified *Anopheles maculipennis* sl. (3 each from Cliffe marshes and Elmley marshes) were DNA-barcoded to determine the species. All were *An. atroparvus* (John Day, unpublished data) and indeed these sites are consistent with the broad environmental preferences of species members of the *An. maculipennis* complex as described in the literature, and with previous records from the North Kent Marshes (Hutchinson, 2004; Snow, 1998; Becker *et al.*, 2010). All recorded *An. maculipennis* s.l. were therefore assumed to be *An. atroparvus* and are referred to as such herein. The other mosquito species present were *Culex pipiens*, *Cx. modestus* and *Culiseta annulata*.

Other ditch fauna were identified morphologically to the most precise level possible in the field (Croft, 1986). Water temperature, salinity and oxidation-reduction potential (ORP) were recorded at each dip site and round using a digital probe (YSI

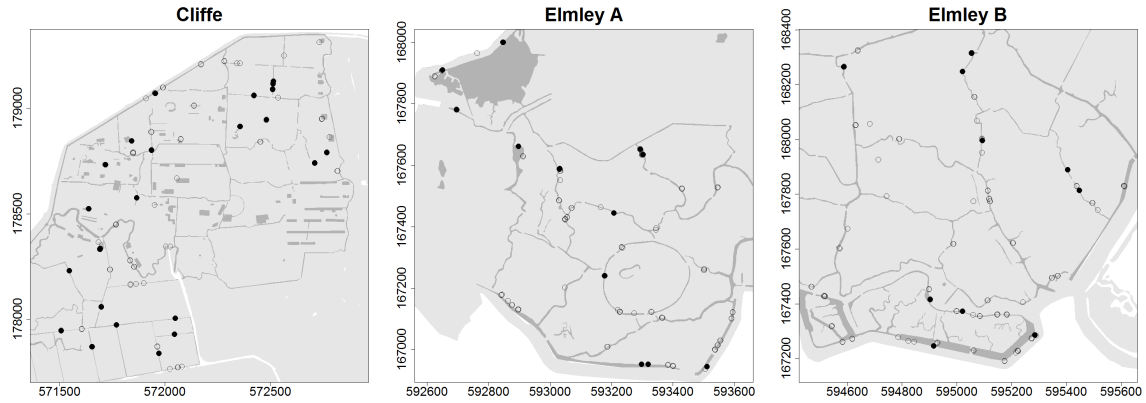


Fig. 3.1: Maps of the three larval survey plots. Filled circles are dip points where mosquitoes were found in at least one dipping round and empty circle where they were not. Grazing marsh is shown in light grey, inland waterbodies (from Ordnance Survey MasterMap) in middle grey and sea in white. Axes are all-numeric Ordnance Survey National Grid references, giving distances in metres.

556 MPS, Yellow Springs, Ohio, USA). Water depth was recorded as the mean of the depth at the edge and the centre of each dip site. High-resolution digital photographs (FinePix XP10, Fujifilm, Minato-ku, Japan) were taken of vegetation at the edge and centre dip points and the presence or absence of different vegetation types at each dipsite was determined from these using field guides (Rose & O'Reilly, 2006; Press & Gibbons, 2002; Rose, 1989). The nine vegetation types recorded are shown in Fig. 3.C.1.

Environmental conditions

To identify potential abiotic environmental drivers of distributions and reduce the number of parameters in the SIDM, we used a stepwise selection procedure to select a subset of abiotic environmental predictors for each faunal taxon. At each stage of the procedure, we parameterized univariate probit regression models by maximum likelihood estimation. Starting from a null model, abiotic environmental predictors were added to the model to select combinations of predictors that minimised the model Akaike Information Criterion (AIC, Akaike (1973)). Fixed terms for dipping rounds and survey area (Elmley or Cliffe Marshes) were included in all of these models

to account for potential correlation structures which could be caused by the repeated-measures study design. The full set environmental covariates from which these subsets were selected consisted of four physical measurements of the water body: depth, temperature, ORP and salinity; and dummy variables indicating presence or absence of the nine vegetation types.

The presence or absence of vegetation types acts as a visible indicator of long-term environmental conditions at dip sites. The plants also strongly influence the structure of the aquatic habitat, variously providing shade or cover from predators. For these reasons we consider vegetation to act as an abiotic environmental condition on mosquito larvae and other aquatic fauna, rather than interacting with them in intimate pair-wise species interactions.

Species interaction distribution model

We use a Bayesian multivariate probit regression model to explicitly model the fundamental niches of each species as well as correlations between the distributions of the different species. Our approach is similar to the model of Ovaskainen *et al.* (2010) but uses the probit function, rather than the logit function as a link. As a result of this modification, we are able to fit the model using a highly efficient Gibbs sampler (Edwards & Allenby, 2003), which greatly speeds up the model fitting process. Full details of the model specification, implementation and choice of priors are given in appendices 3.A & 3.B. We provide software to fit the model as a free, open-source package `BayesComm` (Golding, 2013) for the statistical programming environment R. The package can be downloaded from the CRAN repositories.

Effect Sizes

As well as identifying environmental conditions associated with the presence or absence of mosquitoes, we wish to compare the strength of these effects, to pick out the environmental factors which have the most impact on their distribution. Standardising continuous variables (so that they have a mean of 0 and a standard deviation

of 1) enables us to compare the impact of different environmental factors on species distributions without the confounding effects of different measurement scales. The coefficient can therefore be interpreted as the effect of a 1 standard deviation change in the covariate on the distribution of each mosquito species. Since the discrete variables (such as the presence or absence of a vegetation type) are already on the same scale, these can be compared directly as effect sizes. Unfortunately it is not possible to directly compare the effect size of discrete variables with continuous variables.

Comparing models

As in Ovaskainen *et al.* (2010) we fit four types of model: a *null* model, a *community-only* model, an *environment-only* model and a *full* model. All models were fitted with intercept terms as well as the indicator variables used in the stepwise selection procedure. In the environment-only and full models, for each species the environmental covariates selected by stepwise selection were included in the design matrix. In the null and environment-only models the correlation matrix was set as an identity matrix, enforcing independence of the model errors between species. In the community-only and full models the inter-species correlation matrix was also parameterised.

The null model assumes that each taxon occurs with equal probability at each site (conditional on the dipping round and survey area). Any deviation from this prediction can therefore be interpreted as the distribution of the fauna within the study areas. We quantify these distributions as the residual deviance of the null model. By calculating the residual deviance of the other models and the proportion of the null deviance remaining, we measure the proportion of each species' distribution explained by each model. It should be noted though that since inter-species interactions and environmental covariates are fitted in different ways within the model it is not possible to draw conclusions about the relative importance of the biotic and abiotic factors in driving distributions.

Adding additional parameters to any statistical model inevitably increases its fit to the data, even if there is no true underlying relationship. It is therefore advisable

to account for this potential for overfitting when comparing models. A common approach is to use information criteria which penalize models according to the number of parameters added. Our model includes latent variables and some prior information and an error structure and therefore the number of model parameters is not clearly defined. Spiegelhalter *et al.* (2002) proposed the Deviance Information Criterion (DIC) as a natural approach to compare such models. We therefore calculate DIC and use this to compare the likely predictive power of the different models considered here. Lower DICs indicate better fit with a difference greater than 5 indicating an appreciable difference in explanatory power between models.

Spatial autocorrelation

We checked for a residual spatial autocorrelation structure in each species' distribution. First we calculated raw residuals using the mean probability of presence at each site as predicted by the full model. We split the dataset by dipping round and survey area, resulting in 8 separate sets of residuals for each taxon. We screened these residuals using a Moran's I test in the `spdep` package (Cliff & Ord, 1981; Bivand *et al.*, 2011). For sets with a Moran's I p-value lower than 0.05 we plotted a correlogram using the `ncf` package (Bjornstad, 2012). We visually inspected these correlograms for signs of a coherent spatial autocorrelation structure and detected none.

All models were fitted using `BayesComm` version 0.4 and analysis performed in R version 2.14.2 (R Development Core Team, 2012). The dataset and a full R script to repeat our analysis and reproduce all of the figures in this manuscript are provided in the supplementary material.

Results

The aquatic faunal community we studied contained 16 taxa, including larvae of four mosquito species - all potential vectors of human disease: *Anopheles atroparvus*, a former European malaria vector; *Culex pipiens*, a known maintenance vector of WNV in southern Europe; *Culiseta annulata*, a potential bridge vector of WNV; and *Culex modestus*, a highly efficient bridge vector of WNV only recently found to be breeding in the UK (Becker *et al.*, 2010; Golding *et al.*, 2012). Fig. 3.2 illustrates the observed pattern of co-occurrence between the mosquito species and the other recorded fauna. *Cs. annulata* was only present at Elmley whereas the other three species were present at both study sites.

Environmental drivers

Fig. 3.3 illustrates effect sizes of the AIC-selected environmental covariates for individual mosquito distributions. Estimates of the environmental regression coefficients for all fauna are given in Fig. 3.C.1.

The principle abiotic environmental drivers of the mosquito community distribution were water depth and surface vegetation cover. All species showed a preference for shallow water, with this covariate having the strongest effect on the distribution of *Cs. annulata*. The distributions of *Cx. pipiens*, *Cx. modestus* and *An. atroparvus* were all positively associated with the presence of surface vegetation (filamentous algae, water crowfoot, and duckweed). The distribution of *Cs. annulata* appeared to be unaffected by the presence or absence of any vegetation type, though it was far more likely to be found in cooler water. Salinity had a slight impact on the distributions of two of the mosquito species, with *Cx. modestus* more likely to occur in more saline water and *An. atroparvus* in less saline water.

Environmental covariates explained 20% of spatial variation in the distribution of the entire mosquito community (Fig. 3.5). At the species level, the explained variation ranged from 12% for *Cx. modestus* and *An. atroparvus* to 67% for *Cs. annulata*.

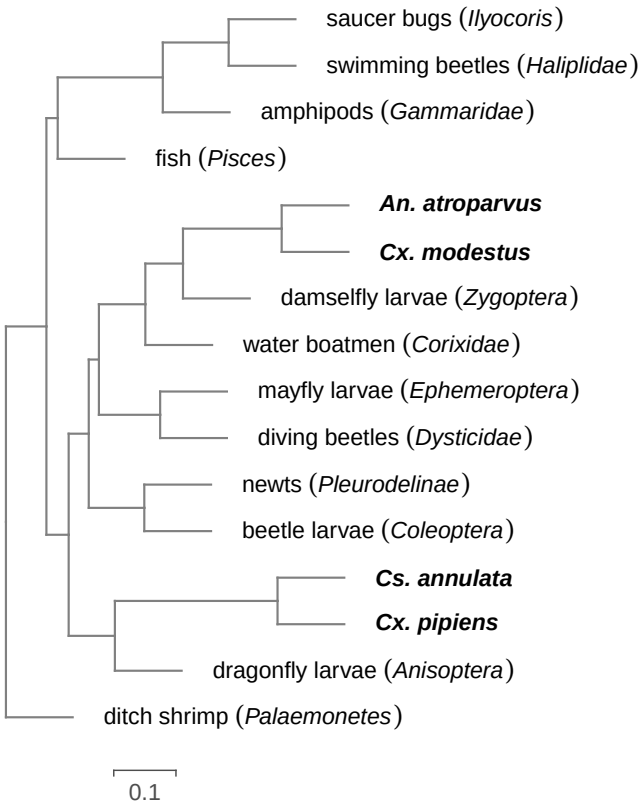


Fig. 3.2: Cluster dendrogram illustrating patterns of co-occurrence between faunal taxa. The dendrogram was created using hierarchical clustering of the raw distribution data. Mosquito species are in boldface and the scale bar gives the dissimilarity along the edges of the dendrogram. Dissimilarity between two species was calculated as one minus the empirical Pearson correlation coefficient with 0 representing perfect positive correlation, 1 no correlation and 2 perfect negative correlation.

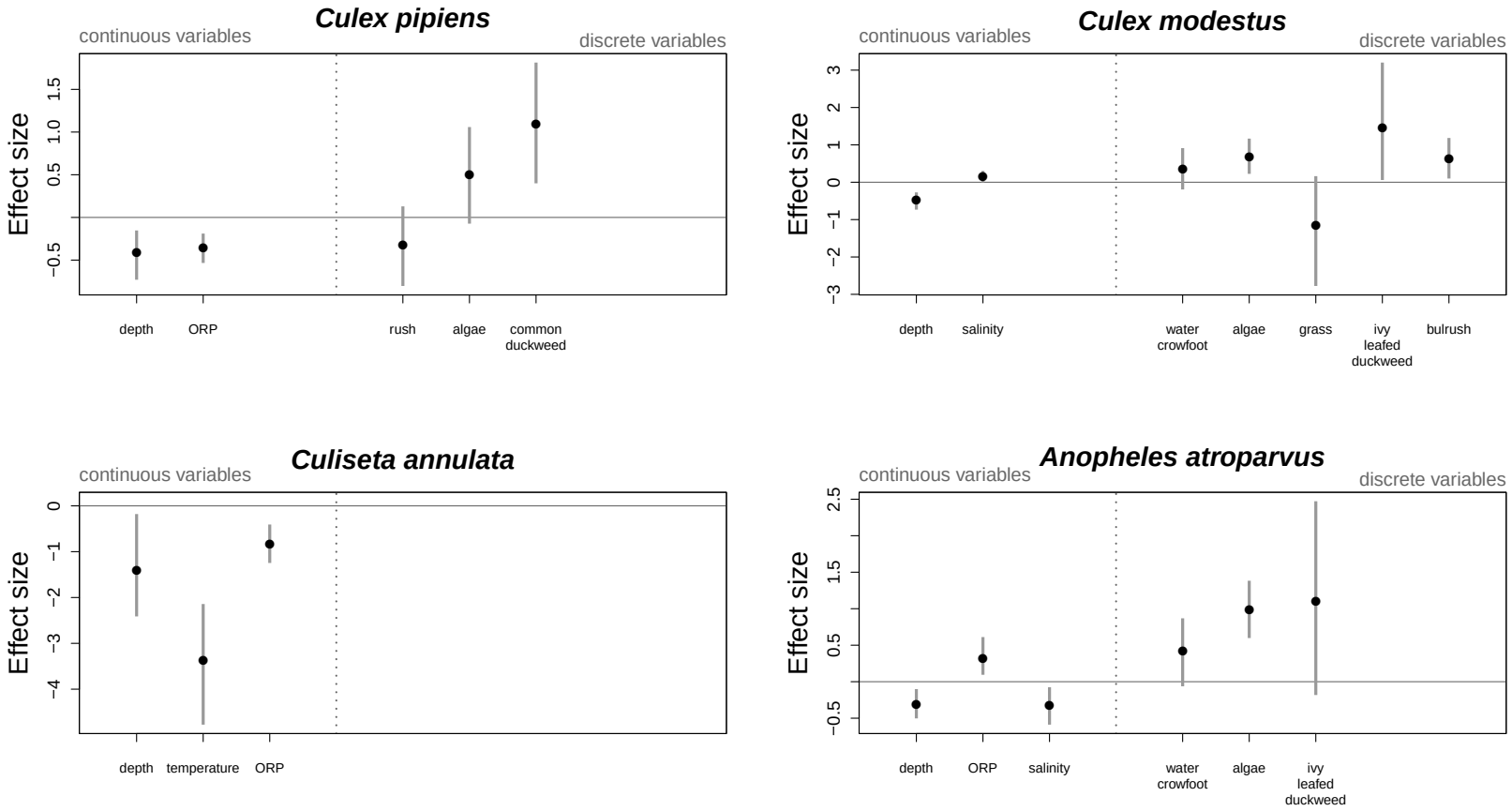


Fig. 3.3: Effect sizes of AIC-selected abiotic environmental covariates used to describe the spatial distribution of the four mosquito species estimated using a full model. Points give the maximum *a posteriori* estimates of effect sizes and grey lines give the associated 95% credible intervals. Credible intervals which do not cross zero signify that the posterior probability that the effect size is 0 or of the opposite sign is less than 0.05, analogous to statistical significance at the 5% level in a frequentist statistical test. Effect sizes are divided into those for continuous variables (depth, temperature, ORP and salinity) and for discrete variables (plant type), since effect sizes cannot be directly compared for the two different types of data.

Biotic interactions

Estimated inter-species correlation coefficients before and after accounting for environmental covariates are illustrated in Fig. 3.4. The parameter estimates and proportion of spatial variation explained are given in Fig. 3.C.2.

The majority of correlation estimates in the full model were positive (67.5% from community-only model and 59% for full model). Correlation coefficients between members of the mosquito community were strongly positive, ranging from 0.28 to 0.49. There were a number of positive correlations between mosquito species and other faunal taxa, including beetle larvae (0.23 to 0.3) and damselfly larvae (0.12 to 0.27). After accounting for each species' fundamental niche, the distributions of both ditch shrimp (*Palaemonetes*) and fish were negatively correlated with those of the four mosquito species (Fig 3.C.2). Correlation coefficients ranged from -0.16 to -0.27 for ditch shrimp and from -0.22 to -0.27 for fish, though the uncertainty around these estimates was larger for *Cx. pipiens* and *Cs. annulata* than for the other two species.

The full model had greater overall explanatory power (after accounting for model complexity) than any of the other models, with a DIC of 3056 (cf. 3648, 3550 and 3150 for the null, community-only and environment-only models).

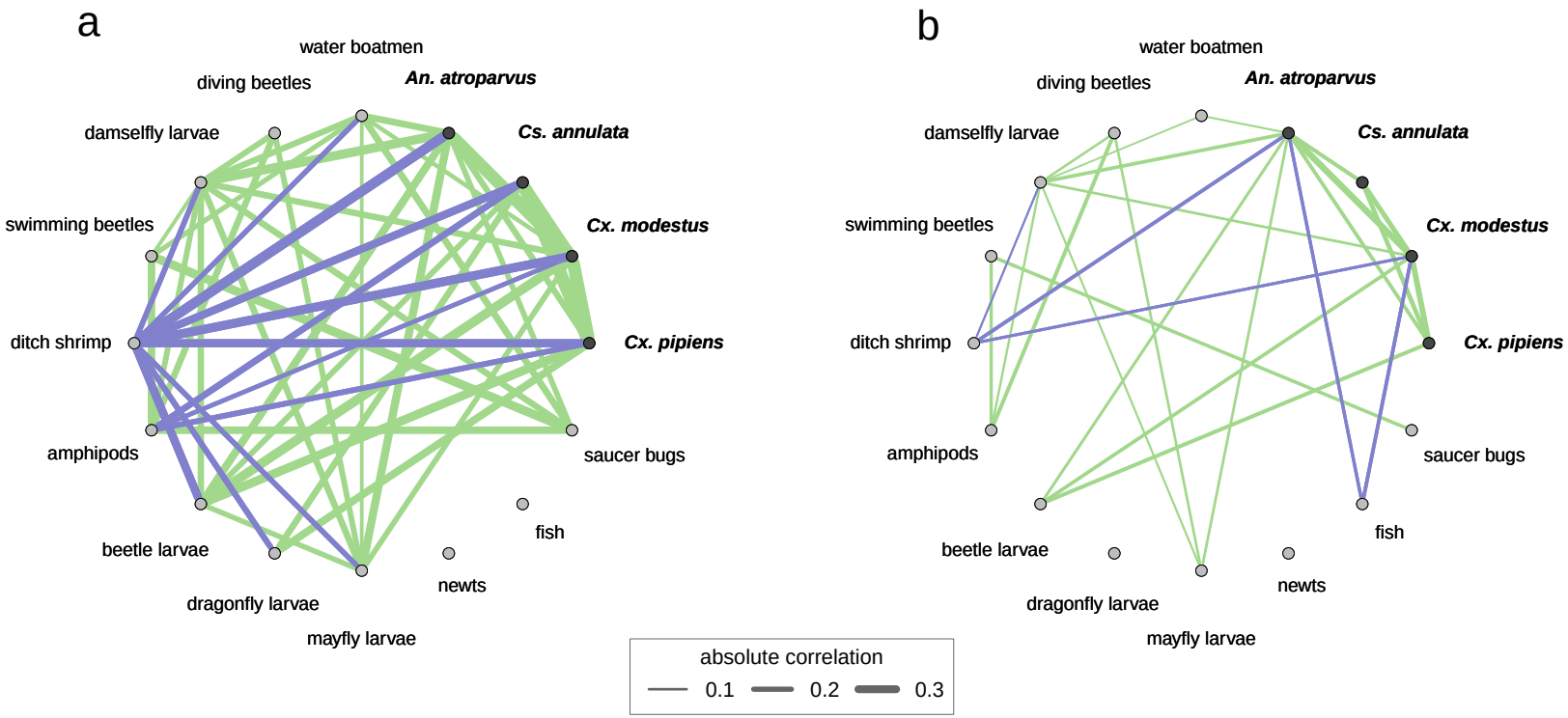


Fig. 3.4: Correlation networks between species in the community a) before accounting for each species' fundamental niche (community-only model) and b) after accounting for fundamental niches (full model). Positive correlations are shown in green and negative in blue. The absolute size of the posterior mean correlation coefficient is reflected in the line width. Correlations are only plotted where the posterior probability of a zero or opposite-sign correlation (Bayesian p-value) was less than 0.05.

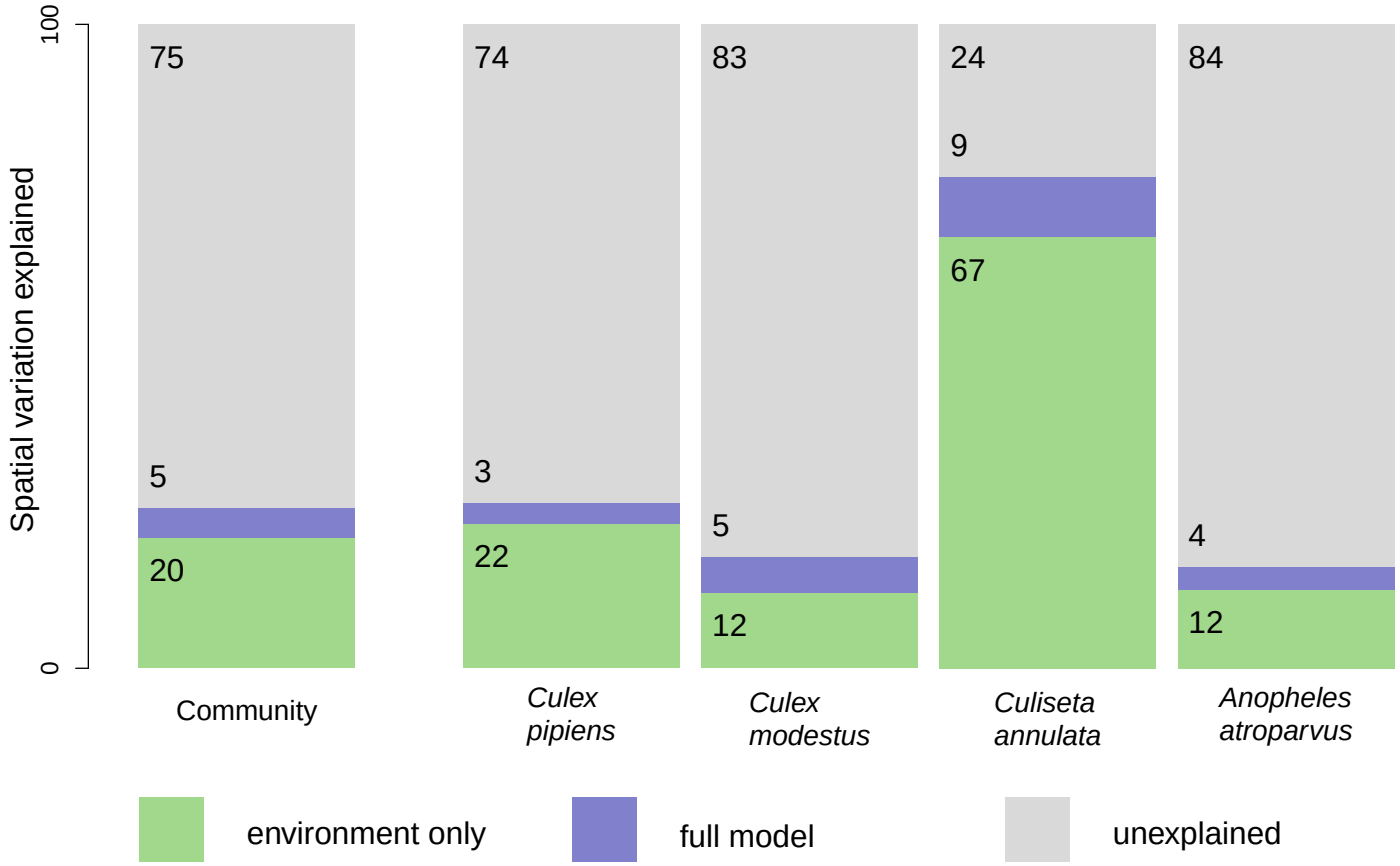


Fig. 3.5: Proportion of spatial variation (measured as deviance) explained by a model with only abiotic environmental covariates and additional variance explained by inclusion of community interactions for the four mosquito species and across the entire mosquito community.

Discussion

Abiotic environmental predictors

The habitat preferences shown by the mosquito species were consistent with existing knowledge of their ecology. All species were more likely to occur in shallower water, with *Cx. pipiens*, *Cx. modestus* and *An. atroparvus* more likely to occur where floating vegetation was present. Such habitats are likely to afford greater protection from predators (Becker *et al.*, 2010)

The positive association between *An. atroparvus* and filamentous algae is in accordance with the studies of Hutchinson (2004) at Elmley marshes: the species was associated with small pools above thick filamentous algae of the genus *Enteromorpha* which were 4°C warmer, less saline and contained fewer predators than sites without algae. Environmental conditions in this microhabitat appear to be advantageous for mosquito larvae, which are highly susceptible to predation and rely on temperature for their development. Unlike the other members of the mosquito community at these sites, which are Culicine, the morphology of Anopheline mosquitoes allows them make use of microhabitats such as these which have limited space (Becker *et al.*, 2010).

The preference of *Cs. annulata* for cooler water and no obvious response to surface vegetation accords with the species' habitat generalism and common use of shaded water bodies. The strong negative effect of ORP on the distribution of *Cs. annulata* and *Cx. pipiens* may reflect their apparent preference for nitrogen rich water (Marshall, 1938), though direct interpretation of ORP alone is difficult.

The negative relationship between the presence of emergent grass (indicative of temporary flooding of normally terrestrial habitat) and the distribution of *Cx. modestus* may signify a preference for more permanent water bodies. *Cx. modestus* was more likely to be found in water beneath bulrushes, which are known to provide a habitat in which adults of the species hibernate (Mouchet *et al.*, 1969).

Biotic interactions

Most of the non-mosquito fauna we recorded in the community (mayfly larvae being the exception) have been incriminated as potential predators of mosquito larvae in the UK (Medlock & Snow, 2008). Wild caught diving beetles, swimming beetles, newts, damselfly larvae and dragonfly larvae have been shown to have consumed larvae (Onyeka, 1983) and amphipods, ditch shrimp, fish, water boatmen and beetle larvae were shown to feed on live larvae in laboratory experiments (Roberts, 1995; Jeffries, 1988; Medlock & Snow, 2008). Of these predators, our model provides evidence only for ditch shrimp and fish being negatively correlated with mosquito larvae after accounting for environmental covariates. Both of these taxa were found to be particularly voracious predators of mosquito larvae under laboratory conditions, consuming in the region of 30 larvae per hour and with ditch shrimp reportedly killing more larvae than they could eat (Roberts, 1995). By comparison, amphipods consumed 1-2 larvae per hour and *Notonecta glauca* nymphs only one larva in 12 hours (Jeffries, 1988). The voracity of ditch shrimp and fish may indicate that their predation, unlike that of other predators, has an appreciable impact on the distribution of mosquito larvae. The identification of fish as effective predators of mosquito larvae is in accordance with findings in other parts of the world, where the introduction of exotic fish or the application of native fish has been successfully used to control mosquito numbers (Becker *et al.*, 2010).

Of the 11 correlation coefficients between mosquito larvae and other fauna which had good evidentiary support (those in Fig. 3.C.2b), 7 were positive and 4 negative. It seems unlikely that such a high proportion of these interactions could represent mutualistic interactions, particularly given the probable predatory nature of the other fauna. A possible explanation is that these positive correlations could indicate low-level predation, with the other fauna being attracted to mosquito prey, but not consuming enough to impact on their distribution (at least measured as presence or absence). A more plausible explanation is that these simply represent shared

responses to environmental variables which have not been completely explained by the model. This would also explain the strong positive correlations between mosquito species, which might more realistically be expected to compete for resources (Juliano, 2009).

Explanatory power

We considered the distribution of mosquito larvae at a very fine spatial resolution, sampled in brief seasonal snapshots. The presence of mosquito larvae at this scale is likely to be influenced by a number of processes other than those considered here, including: the distribution of blood-meal hosts, local population dynamics and dispersal behaviour of larvae and adults. Despite this, we were able to explain 25% of the distribution of the mosquito community (and 76% of the distribution of *Cs. annulata*).

Whilst the distribution of *Cs. annulata* was very well explained in our models, the large parameter estimate for the study area dummy variable in both the environment-only and full models (Fig. 3.C.1) indicates that the species' absence at Cliffe marshes was not explained either by measured environmental covariates or biotic interactions. Why this relatively common species was absent on all visits to this site is as yet unclear.

Advantages and limitations of SIDMs

SIDMs present a promising new method of understanding how interactions within ecological communities can influence species distributions. As with any observational approach to understanding complex systems, SIDMs cannot provide concrete answers to ecological questions. Experimental manipulations of field populations are crucial to establish what factors actually drive community dynamics. However, SIDMs allow the wealth of available observational data on species co-occurrences to be used to refine hypotheses about the drivers of species' distributions (Wisz & Guisan, 2009). SIDMs are likely to be particularly successful at identifying biotic interactions where

other abiotic drivers of species' distributions can be well parameterised and where additional sources of information (such as temporal data) are available (Kissling *et al.*, 2011).

The relative complexity of SIDMs and large number of model parameters makes fitting such models computationally demanding. As a result, development of SIDMs has only recently become feasible for analysing real ecological datasets. By implementing a computationally efficient sampler for multivariate binomial regression and disseminating it as an R package, we hope to contribute to making these approaches available for routine use by ecologists.

Acknowledgements

We thank Steve Gordon at Elmley NNR for his help with fieldwork and acknowledge funding from the NERC Centre for Ecology & Hydrology (Environmental Change Integrating Fund programme).

References

Akaike, H. (1973) Information Theory and an Extension of the Maximum Likelihood Principle. B.N. Petrov & F. Caski, eds., *Proceedings of the Second International Symposium on Information Theory*, pp. 267 – 281. Budapest.

Albert, J.H. & Chib, S. (1993) Bayesian Analysis of Binary and Polychotomous Response Data. **88**, 669–679.

Becker, N., Petric, D., Zgomba, M., Boase, C., Madon, M., Dahl, C. & Kaiser, A. (2010) *Mosquitoes and Their Control*. Springer Verlag, Berlin, second edition.

Beketov, M.A., Yurchenko, Y.A., Belevich, O.E. & Liess, M. (2010) What Environmental Factors are Important Determinants of Structure, Species Richness, and Abundance of Mosquito Assemblages? *Journal of Medical Entomology*, **47**, 129–139.

Bivand, R., with contributions by Micah Altman, Anselin, L., Assuno, R., Berke, O., Bernat, A., Blanchet, G., Blankmeyer, E., Carvalho, M., Christensen, B., Chun, Y., Dormann, C., Dray, S., Halbersma, R., Krainski, E., Legendre, P., Lewin-Koh, N., Li, H., Ma, J., Millo, G., Mueller, W., Ono, H., Peres-Neto, P., Piras, G., Reder, M., Tiefelsdorf, M. & Yu., D. (2011) *spdep: Spatial dependence: weighting schemes, statistics and models*. R package version 0.5-43.

Bjornstad, O.N. (2012) *ncf: spatial nonparametric covariance functions*. R package version 1.1-4.

Blaustein, L. & Chase, J.M. (2007) Interactions between mosquito larvae and species that share the same trophic level. *Annual Review of Entomology*, **52**, 489–507.

Brown, H.E., Diuk-Wasser, M.A., Andreadis, T.G. & Fish, D. (2008) Remotely-sensed vegetation indices identify mosquito clusters of West Nile virus vectors in an urban landscape in the northeastern United States. *Vector-Borne and Zoonotic Diseases*, **8**, 197–206.

Cliff, A. & Ord, J. (1981) *Spatial Processes: Models & Applications*. Pion.

Cranston, P., Ramsdale, C.D., Snow, K.R. & White, G. (1987) *Adults, Larvae, and Pupae of British Mosquitoes (Culicidae) A Key*. Freshwater Biological Association.

Croft, P. (1986) *A key to the major groups of British freshwater invertebrates*, volume 6. Field Studies Council.

Diuk-Wasser, M.A., Brown, H.E., Andreadis, T.G. & Fish, D. (2006) Modeling the Spatial Distribution of Mosquito Vectors for West Nile Virus in Connecticut, USA. *Vector-Borne and Zoonotic Diseases*, **6**, 283–295.

Durand, B., Balança, G., Baldet, T. & Chevalier, V. (2010) A metapopulation model to simulate West Nile virus circulation in Western Africa, Southern Europe and the Mediterranean basin. *Veterinary Research*, **41**, 32.

Edwards, Y. & Allenby, G. (2003) Multivariate analysis of multiple response data. *Journal of Marketing Research*, **40**, 321–334.

Ferguson, H.M., Dornhaus, A., Beeche, A., Borgemeister, C., Gottlieb, M., Mulla, M.S., Gimnig, J.E., Fish, D. & Killeen, G.F. (2010) Ecology: A Prerequisite for Malaria Elimination and Eradication. *PLoS Medicine*, **7**, e1000303.

Golding, N. (2013) *BayesComm: Bayesian community ecology analysis*. R package version 0.1-0.

Golding, N., Nunn, M.A., Medlock, J.M., Purse, B.V., Vaux, A.G.C. & Schäfer, S.M. (2012) West Nile virus vector Culex modestus established in southern England. *Parasites & Vectors*, **5**, 32.

Hay, S.I., Sinka, M.E., Okara, R.M., Kabaria, C.W., Mbithi, P.M., Tago, C.C., Benz, D., Gething, P.W., Howes, R.E., Patil, A.P., Temperley, W.H., Bangs, M.J., Chareonviriyaphap, T., Elyazar, I.R.F., Harbach, R.E., Hemingway, J., Manguin, S.,

Mbogo, C.M., Rubio-Palis, Y. & Godfray, H.C.J. (2010) Developing global maps of the dominant anopheles vectors of human malaria. *PLoS Medicine*, **7**, e1000209.

Hutchinson, R.A. (2004) *Mosquito Borne Diseases in England: past, present and future risks, with special reference to malaria in the Kent Marshes.* Ph.D. thesis, Durham.

Jeffries, M. (1988) Individual vulnerability to predation: the effect of alternative prey types. *Freshwater Biology*, **19**, 49–56.

Juliano, S.A. (2009) Species interactions among larval mosquitoes: context dependence across habitat gradients. *Annual Review of Entomology*, **54**, 37–56.

Kissling, W.D., Dormann, C.F., Groeneveld, J., Hickler, T., Kühn, I., McInerny, G.J., Montoya, J.M., Römermann, C., Schiffers, K., Schurr, F.M., Singer, A., Svenning, J.C., Zimmermann, N.E. & OHara, R.B. (2011) Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *Journal of Biogeography*, **39**, 2163–2178.

Marshall, J.F. (1938) *The British Mosquitoes.* Trustees of the British Museum.

Medlock, J.M. & Snow, K.R. (2008) Natural predators and parasites of British mosquitoes a review. *European Mosquito Bulletin*, **25**, 1–11.

Mouchet, J., Rageau, J. & Chippaux, A. (1969) Hibernation de Culex modestus Ficalbi (Diptera, Culicidae) en Camargue. *Cahiers ORSTOM Serie Entomologie Medicale et Parasitologie*, **7**, 35–37.

Mukabana, W.R., Kannady, K., Kiama, G.M., Ijumba, J.N., Mathenge, E.M., Kiche, I., Nkwengulila, G., Mboera, L., Mtasiwa, D., Yamagata, Y., van Schayk, I., Knols, B.G.J., Lindsay, S.W., Caldas de Castro, M., Mshinda, H., Tanner, M., Fillinger, U. & Killeen, G.F. (2006) Ecologists can enable communities to implement malaria vector control in Africa. *Malaria Journal*, **5**, 9.

Onyeka, J.O.A. (1983) Studies on the natural predators of Culex pipiens L. and C. torrentium Martini (Diptera: Culicidae) in England. *Bulletin of Entomological Research*, **73**, 185–194.

Ovaskainen, O., Hottola, J. & Siitonen, J. (2010) Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology*, **91**, 2514–21.

Press, J. & Gibbons, B. (2002) *Wild Flowers of Britain and Europe*. Photographic Field Guide Series. New Holland.

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.

Reisen, W.K. (2010) Landscape epidemiology of vector-borne diseases. *Annual Review of Entomology*, **55**, 461–83.

Roberts, G. (1995) Salt-marsh crustaceans, Gammarus duebeni and Palaemonetes varians as predators of mosquito larvae and their reaction to Bacillus thuringiensis subsp. israelensis. *Biocontrol Science and Technology*, pp. 37–41.

Rose, F. (1989) *Colour identification guide to the grasses, sedges, rushes and ferns of the British Isles and north-western Europe*. Viking, London.

Rose, F. & O'Reilly, C. (2006) *The Wild Flower Key*. Warne, London.

Schäfer, M.L., Lundström, J.O., Pfeffer, M., Lundkvist, E. & Landin, J. (2004) Biological diversity versus risk for mosquito nuisance and disease transmission in constructed wetlands in southern Sweden. *Medical and Veterinary Entomology*, **18**, 256–67.

Sinka, M.E., Bangs, M.J., Manguin, S., Coetzee, M., Mbogo, C.M., Hemingway, J., Patil, A.P., Temperley, W.H., Gething, P.W., Kabaria, C.W., Okara, R.M., Van

Boeckel, T.P., Godfray, H.C.J., Harbach, R.E. & Hay, S.I. (2010) The dominant Anopheles vectors of human malaria in Africa, Europe and the Middle East: occurrence data, distribution maps and bionomic precis. *Parasites & Vectors*, **3**, 117.

Snow, K.R. (1990) *Mosquitoes (Naturalists' Handbooks 14).* Richmond Publishing Company.

Snow, K.R. (1998) Distribution of Anopheles mosquitoes in the British Isles. *European Mosquito Bulletin*, **1**, 9–13.

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**, 583–639.

Steiger, D., Johnson, P., Hilbert, D., Ritchie, S., Jones, D., Laurance, S. & Laurance, G. (2012) Effects of landscape disturbance on mosquito community composition in tropical Australia. *Journal of Vector Ecology*, **37**, 69–76.

Tran, A., Ponçon, N., Toty, C., Linard, C., Guis, H., Ferré, J.B., Lo Seen, D., Roger, F., de la Rocque, S., Fontenille, D. & Baldet, T. (2008) Using remote sensing to map larval and adult populations of Anopheles hyrcanus (Diptera: Culicidae) a potential malaria vector in Southern France. *International Journal of Health Geographics*, **7**, 9.

Wisz, M.S. & Guisan, A. (2009) Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. *BMC Ecology*, **9**, 8.

Wootton, J.T. & Emmerson, M. (2005) Measurement of Interaction Strength in Nature. *Annual Review of Ecology, Evolution, and Systematics*, **36**, 419–444.

Appendix 3.A Statistical model and inference

3.A.1 Statistical model

We use a multivariate extension of the latent variable model for binary regression (Albert & Chib, 1993). Our approach is very similar to a model recently described for analysis of ecological communities (Ovaskainen *et al.*, 2010) except that we draw the latent variable from a normal, rather than a logistic, distribution. This is equivalent to using a probit function rather than a logit function as the canonical link in a univariate binomial regression. The probit and logit functions are very similar, with the only apparent drawback of the probit model being that regression coefficients for binary covariates can no longer be interpreted directly as log odds ratios (which are not widely used in ecology). The advantage of our approach is that it enables use of a Gibbs sampler to make inference about the regression parameters, thereby reducing computation time and the risk of numerical errors. The model is defined as:

$$\begin{aligned} y_{ij} &= 1(z_{ij} > 0) \\ z_{ij} &= \mu_{ij} + e_{ij} \\ \mu_{ij} &= \mathbf{X}_j \boldsymbol{\beta}_j \\ \mathbf{e}_i &\sim N(\mathbf{0}, \mathbf{R}) \end{aligned} \tag{3.A.1}$$

where $y_{i,j}$ is a binomial variable representing presence (1) or absence (0) of species j at site i, z is a normally distributed latent variable, $1(z > 0)$ is an indicator function returning 1 when $z > 0$ and 0 otherwise, $\mathbf{X}_j$ is an n by k_j design matrix for species j, $\boldsymbol{\beta}_j$ is a vector of k_j regression coefficients for species j and $N(\mathbf{0}, \mathbf{R})$ is an m-dimensional standard multivariate normal distribution with mean vector $\mathbf{0}$ and symmetric, positive-definite correlation matrix $\mathbf{R}$, n is the number of sites, m the number of species and k_j the number of environmental covariates used to model the fundamental niche of species j. The elements of $\mathbf{R}$ describe whether species co-occur more or less often than would be expected by their fundamental niches alone and is indicative of the underlying network of interactions between species in the community.

3.A.2 Model inference

We use a computationally efficient Gibbs algorithm, based on a sampler described by Edwards & Allenby (2003), to sample the model parameters in turn from their conditional posterior distributions. At each iteration this entails the following steps:

1. Sample the latent variables $\mathbf{z}$ from a truncated multivariate standard normal distribution:

$$\mathbf{z} \sim N_T(\boldsymbol{\mu}, \mathbf{R}) \tag{3.A.2}$$

such that z_{ij} is positive when $y_{ij} = 1$ and negative otherwise.

2. Sample the vector of regression coefficients $\boldsymbol{\beta}_j$ for each species j from a multivariate normal distribution:

$$\boldsymbol{\beta}_j \sim N((\sigma\mathbf{I} + \mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j'\mathbf{z}_j, (\sigma\mathbf{I} + \mathbf{X}_j'\mathbf{X}_j)^{-1}) \tag{3.A.3}$$

where σ is the standard deviation of the prior distribution over $\boldsymbol{\beta}_j$, $\mathbf{I}$ is an identity matrix, having diagonal elements 1 and all other elements 0 and $'$ denotes the matrix transpose.

3. Sample the correlation matrix $\mathbf{R}$ by first sampling a covariance matrix $\mathbf{W}$ and scaling this to a correlation matrix:

$$\begin{aligned} \mathbf{W}^{-1} &\sim \mathcal{W}_m(\nu, \mathbf{e}'\mathbf{e} + \mathbf{S}) \\ \mathbf{C} &= diag(\mathbf{W})^{-\frac{1}{2}} \\ \mathbf{R} &= \mathbf{CWC}' \end{aligned} \tag{3.A.4}$$

where $diag(.)$ denotes the diagonal vector of a matrix and $\mathcal{W}_m(.,\ .)$ is a Wishart distribution of dimension m (the number of species). The scale matrix $\mathbf{S}$ and degrees of freedom parameter ν define the prior distribution over $\mathbf{W}$ and therefore $\mathbf{R}$.

By sampling a covariance matrix and scaling to a correlation matrix, we are able to use a Gibbs sampler whilst avoiding the non-identifiability of the variance of a model for binary data. The implications of this approach are discussed in Edwards & Allenby (2003). Our choices of prior parameters σ, $\mathbf{S}$ and ν are discussed in Appendix 3.B

Appendix 3.B Choice of priors

To construct a sampler for Bayesian statistical models it is necessary to specify priors over all of the parameters for which we want to make inference. These express our prior belief (before observing any data) about the probability distribution of the parameters. Here we use conjugate priors to enable the efficient Gibbs sampler described above.

For each element of each vector of regression coefficients $\boldsymbol{\beta}_j$ we use a diffuse normal prior with mean 0 and variance 100 by setting σ to 10. This is a widely used prior which exhibits little influence on the posterior. Specification of an appropriate prior for the unidentified covariance matrix $\mathbf{W}$ is less straightforward. A commonly used conjugate prior for $\mathbf{W}$ is obtained by setting:

$$\begin{aligned} \nu &= m + 1 \\ \mathbf{S} &= \nu \mathbf{I} \end{aligned} \tag{3.B.1}$$

This prior has the feature that each element of a correlation matrix derived from $\mathbf{W}$ has a marginally uniform distribution and it therefore has no impact on the posterior. Such a prior is problematic for our model for two reasons. Firstly, a uniform prior implies that it is equally likely for the distributions of two species to be very strongly correlated (i.e. always found together or never found together) as it is for there to be no correlation between them. This is biologically unrealistic; we would expect the majority of inter-species interactions to be weak or non-existent with relatively few interactions driving moderate correlations in distributions (Wootton & Emmerson, 2005). Secondly, a prior of this sort exhibits a dependency between the unobserved variance parameters of $\mathbf{W}$ and the correlation coefficients of $\mathbf{R}$, such that when these variance parameters are large, the prior assigns much higher probability to strong correlations than weak correlations. This leads to very unrealistic posterior parameter estimates with a bimodal distribution close to 1 and -1. Edwards & Allenby (2003) demonstrate a weakly informative prior which avoids this problem but

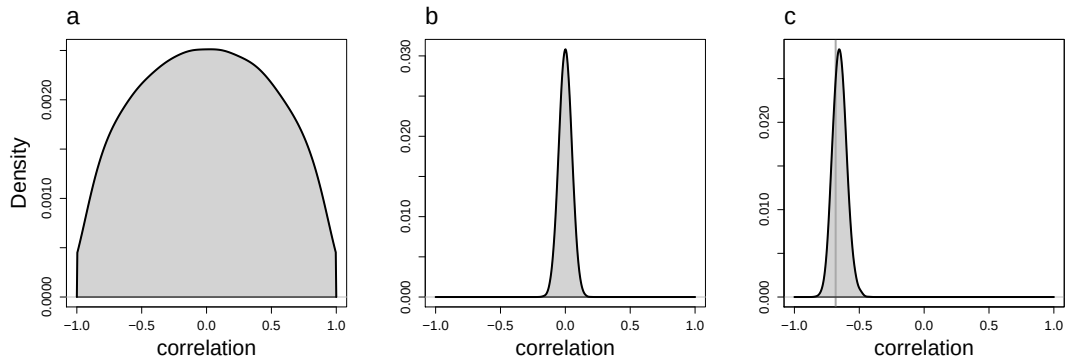


Fig. 3.B.1: Illustration of our inverse Wishart prior over the correlation matrix and its impact on the posteriors. Prior probability density over each element of **R** for a) 100 records and b) 400 records, estimated from 100,000 simulations. c) Posterior probability density of the correlation coefficient from a `BayesComm` model applied to a simulated bivariate dataset with 400 observations. The vertical line shows the maximum-likelihood estimate of the coefficient.

maintains conjugacy for the likelihood:

$$\begin{aligned} \nu &= n + 2m \\ \mathbf{S} &= 2m\mathbf{I} \end{aligned} \tag{3.B.2}$$

where n is the number of observations. With few observations this gives a reasonable non-uniform prior. As the number of observations increases the prior becomes centred on 0, but with the increased amount of data its influence on the posterior becomes weaker. The shape of this prior and its weak, but informative, effect on the posterior are illustrated using simulated data in Fig. 3.B.1 and R code to reproduce these figures is given in the supplementary material.

Appendix 3.C Parameter estimates

The posterior distributions of the regression coefficients from the environment-only and full model are summarised in Fig. 3.C.1. Summaries of the posterior distributions of the correlation coefficients and the percentage of deviance explained per species from the community-only and the full model are shown in Fig. 3.C.2.

a

	Intercept	depth	temperature	ORP	salinity	water crowfoot (Ranunculus)	rushes (Juncus/Scirpus)	filamentous algae	emergent grass	Ivy-leafed duckweed (Lemna trisulca)	bulrushes (Typha)	reeds (Phragmites)	marestail (Hippuris)	common duckweed (Lemna minor)	August 2010	July 2011	August 2011	Cliffe marshes
Cx. pipiens	−1.18 (0.21)	−0.36 (0.12)		−0.33 (0.08)			−0.38 (0.19)	0.4 (0.23)					0.96 (0.3)	−0.1 (0.2)	−0.74 (0.27)	−0.42 (0.27)	0.07 (0.21)	
Cx. modestus	−0.94 (0.14)	−0.43 (0.1)			0.15 (0.07)	0.38 (0.24)		0.57 (0.2)	−1.09 (0.64)	1.3 (0.75)	0.63 (0.23)		0.76 (0.33)		−0.47 (0.17)	−1.27 (0.24)	−0.91 (0.23)	0.15 (0.19)
Cs. annulata	−3.99 (0.7)	−0.77 (0.4)	−1.15 (0.36)	−0.64 (0.16)											−0.03 (0.61)	−0.14 (0.73)	−0.1 (0.98)	−3.25 (1.63)
An. atroparvus	−1.47 (0.16)	−0.26 (0.09)		0.28 (0.12)	−0.35 (0.12)	0.38 (0.23)		0.88 (0.19)		0.98 (0.64)					0.14 (0.2)	−0.55 (0.24)	−0.28 (0.26)	0.11 (0.18)
water boatmen	−0.46 (0.12)	−0.28 (0.06)			−0.23 (0.08)	0.6 (0.2)		0.42 (0.16)		1.35 (0.74)	−0.97 (0.24)	0.33 (0.17)			−0.22 (0.16)	−0.24 (0.16)	−0.55 (0.18)	0.34 (0.15)
diving beetles	−0.92 (0.13)			−0.31 (0.12)											−0.28 (0.18)	−3.63 (1.48)	−3.42 (1.41)	−0.32 (0.2)
damselfly larvae	−1.17 (0.17)					0.6 (0.21)	0.37 (0.15)	0.55 (0.16)	−0.92 (0.63)				0.72 (0.3)		0.33 (0.16)	0.03 (0.16)	−0.05 (0.18)	0.01 (0.14)
swimming beetles	−2.89 (0.39)				−0.27 (0.12)		0.53 (0.24)	0.6 (0.21)	1.29 (0.69)			0.66 (0.25)		0.64 (0.32)	0.6 (0.33)	1.48 (0.3)	1.55 (0.32)	−0.65 (0.22)
ditch shrimp	−0.25 (0.13)			0.41 (0.18)	0.16 (0.06)	−1.31 (0.51)		−1.05 (0.37)			−3.07 (1.45)				0.15 (0.19)	−0.47 (0.24)	−0.47 (0.28)	−1.61 (0.28)
amphipods	−1.04 (0.16)	−0.14 (0.09)		0.25 (0.14)				0.39 (0.25)			−3.17 (1.64)				0.01 (0.22)	0.26 (0.23)	0.16 (0.25)	−1.43 (0.27)
beetle larvae	−3.45 (0.63)	−0.39 (0.17)	0.46 (0.14)				1.42 (0.6)		1.21 (0.66)			−2.71 (1.81)		1.35 (0.49)	−0.55 (0.4)	−0.12 (0.3)	−0.74 (0.47)	0.47 (0.25)
dragonfly larvae	−2.16 (0.32)		−0.32 (0.18)	−0.37 (0.11)											−0.86 (0.46)	0.01 (0.39)	−0.73 (0.58)	−0.06 (0.34)
mayfly larvae	−1.9 (0.28)	−0.31 (0.11)			−0.62 (0.17)		0.43 (0.24)	0.6 (0.22)	−2.7 (1.9)				0.67 (0.39)		−0.13 (0.22)	−0.31 (0.23)	−0.29 (0.29)	−0.39 (0.22)
newts	−5.25 (1.33)	−0.82 (0.38)												1.35 (0.59)	−2.4 (1.31)	−0.07 (0.5)	−0.56 (0.71)	3.04 (1.3)
fish	−2.11 (0.31)	−0.4 (0.18)	−0.38 (0.17)			0.83 (0.34)								−2.99 (2.02)	−1.31 (0.54)	−0.35 (0.39)	−0.51 (0.42)	0.55 (0.29)
saucer bugs	−4.4 (0.92)				−0.3 (0.19)	0.73 (0.31)		1.3 (0.29)			−3.73 (2.08)			1.07 (0.61)	−1.79 (1.63)	2.2 (0.9)	2.62 (0.9)	−0.07 (0.28)

b

Cx. pipiens	−1.26 (0.25)	−0.41 (0.15)		−0.35 (0.09)			−0.33 (0.23)	0.5 (0.28)					1.09 (0.35)	−0.17 (0.23)	−0.81 (0.32)	−0.53 (0.34)	0.05 (0.25)	
Cx. modestus	−1.01 (0.16)	−0.49 (0.12)			0.15 (0.08)	0.36 (0.28)		0.69 (0.24)	−1.14 (0.76)	1.46 (0.78)	0.63 (0.27)		0.81 (0.38)		−0.53 (0.22)	−1.32 (0.28)	−0.95 (0.28)	0.14 (0.22)
Cs. annulata	−6.96 (0.92)	−1.41 (0.54)	−3.36 (0.64)	−0.83 (0.21)											−0.92 (0.77)	−1.05 (0.81)	−0.91 (1.55)	−3.36 (0.95)
An. atroparvus	−1.56 (0.19)	−0.31 (0.11)		0.32 (0.13)	−0.33 (0.13)	0.42 (0.24)		0.98 (0.2)		1.1 (0.69)					0.15 (0.23)	−0.52 (0.26)	−0.29 (0.3)	0.08 (0.2)
water boatmen	−0.46 (0.13)	−0.28 (0.06)			−0.21 (0.09)	0.6 (0.2)		0.44 (0.17)		1.45 (0.84)	−1 (0.28)	0.32 (0.17)			−0.22 (0.16)	−0.24 (0.16)	−0.56 (0.19)	0.33 (0.15)
diving beetles	−0.9 (0.14)			−0.26 (0.14)											−0.28 (0.2)	−3.51 (1.34)	−2.98 (1.12)	−0.34 (0.22)
damselfly larvae	−1.23 (0.19)					0.62 (0.2)	0.43 (0.16)	0.56 (0.16)	−1.12 (0.71)				0.77 (0.32)		0.34 (0.17)	0.05 (0.18)	−0.05 (0.19)	0 (0.14)
swimming beetles	−3.04 (0.45)				−0.25 (0.14)		0.65 (0.29)	0.6 (0.26)	1.43 (0.74)			0.74 (0.28)		0.78 (0.34)	0.61 (0.37)	1.51 (0.33)	1.57 (0.35)	−0.66 (0.25)
ditch shrimp	−0.26 (0.14)			0.37 (0.2)	0.14 (0.06)	−1.02 (0.51)		−1.11 (0.44)			−2.46 (1.43)				0.17 (0.21)	−0.45 (0.25)	−0.47 (0.3)	−1.59 (0.31)
amphipods	−1.02 (0.17)	−0.19 (0.1)		0.26 (0.17)				0.37 (0.26)			−5.24 (2.52)				−0.03 (0.24)	0.21 (0.25)	0.13 (0.27)	−1.48 (0.31)
beetle larvae	−4.71 (1.11)	−0.53 (0.2)	0.35 (0.16)				2.7 (1.06)		1.62 (0.73)			−0.96 (1.54)		1.45 (0.52)	−0.8 (0.46)	−0.27 (0.34)	−0.99 (0.55)	0.53 (0.27)
dragonfly larvae	−2.13 (0.33)		−0.35 (0.2)	−0.36 (0.11)											−0.98 (0.54)	−0.04 (0.42)	−0.77 (0.66)	−0.09 (0.36)
mayfly larvae	−2.13 (0.33)	−0.39 (0.12)			−0.75 (0.2)		0.55 (0.27)	0.66 (0.23)	−6 (2.95)				0.77 (0.42)		−0.14 (0.24)	−0.29 (0.25)	−0.2 (0.32)	−0.39 (0.24)
newts	−4.99 (1.12)	−1.05 (0.65)												1.51 (0.66)	−2.75 (1.47)	−0.18 (0.56)	−0.59 (0.76)	2.63 (0.91)
fish	−2.21 (0.36)	−0.4 (0.2)	−0.26 (0.17)			0.76 (0.34)								−6.32 (2.58)	−1.12 (0.53)	−0.13 (0.43)	−0.36 (0.45)	0.54 (0.3)
saucer bugs	−4.36 (1.01)				−0.32 (0.22)	0.87 (0.32)		1.31 (0.32)			−7.61 (3.47)			1.38 (0.6)	−2.36 (2.2)	2.05 (0.99)	2.48 (1)	−0.02 (0.33)

Fig. 3.C.1: Regression coefficients from a) the environment-only model and b) the full model. Displayed are the posterior means with standard deviations in parentheses. Positive coefficients are in green and negative coefficients in blue. Squares are left blank where coefficients were not selected by AIC.

a

	Cx. pipiens	*Cx. modestus*	*Cs. annulata*	*An. atroparvus*	water boatmen	diving beetles	damselfly larvae	swimming beetles	ditch shrimp	amphipods	beetle larvae	dragonfly larvae	mayfly larvae	newts	fish	saucer bugs
Cx. pipiens	4%	0.49	0.5	0.23	0.05	0	0.09	−0.04	−0.31	−0.21	0.34	0.29	0.06	0.08	−0.13	−0.07
Cx. modestus	0.07	3%	0.36	0.48	0.17	−0.02	0.24	−0.02	−0.34	−0.17	0.36	0.21	0.21	0.14	−0.18	0.07
Cs. annulata	0.09	0.11	7%	0.18	0.03	−0.03	0.02	−0.05	−0.3	−0.25	0.22	0.25	0.07	0.07	−0.1	−0.07
An. atroparvus	0.08	0.07	0.12	2%	0.26	0.06	0.35	0.01	−0.37	0.02	0.3	0.06	0.31	0.12	−0.15	0.21
water boatmen	0.08	0.08	0.11	0.07	0%	0.09	0.24	0.19	−0.19	0.12	0.13	0.07	0.14	0.05	−0.01	0.26
diving beetles	0.1	0.1	0.13	0.1	0.09	0%	0.22	0.07	−0.05	0.25	0.13	−0.02	0.22	0.07	0.04	0.14
damselfly larvae	0.08	0.07	0.12	0.07	0.07	0.09	0%	0.14	−0.21	0.22	0.24	0.01	0.25	0.07	−0.02	0.24
swimming beetles	0.11	0.1	0.13	0.1	0.08	0.13	0.09	1%	−0.09	0.29	0.11	0	−0.06	0.02	0	0.34
ditch shrimp	0.09	0.09	0.11	0.09	0.07	0.1	0.08	0.1	1%	0.11	−0.31	−0.22	−0.2	−0.12	0.06	−0.16
amphipods	0.11	0.1	0.13	0.1	0.08	0.1	0.08	0.09	0.09	2%	−0.01	−0.19	0.03	0	0.05	0.29
beetle larvae	0.1	0.09	0.12	0.1	0.1	0.11	0.09	0.12	0.11	0.12	2%	0.11	0.2	0.17	−0.17	0.09
dragonfly larvae	0.11	0.13	0.13	0.14	0.12	0.14	0.12	0.13	0.13	0.14	0.14	0%	0.02	0.02	0.01	−0.06
mayfly larvae	0.1	0.09	0.13	0.08	0.08	0.1	0.08	0.12	0.1	0.1	0.1	0.13	0%	0.06	−0.12	0.12
newts	0.19	0.17	0.19	0.16	0.14	0.14	0.14	0.15	0.16	0.17	0.16	0.17	0.16	0%	−0.09	0.07
fish	0.16	0.15	0.17	0.14	0.11	0.13	0.12	0.14	0.14	0.14	0.14	0.16	0.14	0.16	0%	0.02
saucer bugs	0.14	0.13	0.15	0.11	0.1	0.14	0.1	0.09	0.13	0.11	0.13	0.16	0.12	0.16	0.14	0%

b

	Cx. pipiens	*Cx. modestus*	*Cs. annulata*	*An. atroparvus*	water boatmen	diving beetles	damselfly larvae	swimming beetles	ditch shrimp	amphipods	beetle larvae	dragonfly larvae	mayfly larvae	newts	fish	saucer bugs
Cx. pipiens	26%	0.43	0.35	0.28	0.05	−0.1	0.14	−0.1	−0.16	−0.1	0.3	0.06	0	0.03	−0.22	−0.14
Cx. modestus	0.08	17%	0.37	0.49	0.11	−0.09	0.21	−0.11	−0.23	−0.15	0.28	0.07	0.15	0.12	−0.27	−0.06
Cs. annulata	0.12	0.12	76%	0.3	0.1	−0.13	0.12	−0.08	−0.2	−0.13	0.23	0.02	0.1	0.06	−0.22	−0.07
An. atroparvus	0.09	0.07	0.14	16%	0.15	0	0.27	−0.13	−0.27	−0.02	0.23	0.02	0.2	0.12	−0.25	−0.01
water boatmen	0.09	0.08	0.12	0.08	7%	0.07	0.14	0.1	−0.13	0.08	0.07	0.09	0.05	0.03	−0.09	0.09
diving beetles	0.11	0.09	0.16	0.1	0.09	2%	0.19	0.05	0.08	0.3	0.07	−0.06	0.2	0.03	0.03	0.1
damselfly larvae	0.09	0.08	0.14	0.08	0.07	0.1	5%	0.06	−0.15	0.18	0.11	0.03	0.17	0	−0.07	0.06
swimming beetles	0.12	0.11	0.15	0.11	0.09	0.13	0.09	10%	−0.02	0.25	−0.05	−0.02	−0.16	−0.07	0.03	0.26
ditch shrimp	0.11	0.1	0.15	0.12	0.08	0.1	0.08	0.1	13%	0.08	−0.15	−0.09	−0.04	−0.05	0.12	−0.04
amphipods	0.11	0.11	0.17	0.11	0.08	0.1	0.08	0.09	0.09	10%	−0.05	−0.08	−0.01	−0.05	0.06	0.18
beetle larvae	0.1	0.1	0.14	0.1	0.1	0.13	0.1	0.14	0.13	0.13	28%	0.02	0.08	0.09	−0.2	−0.07
dragonfly larvae	0.15	0.16	0.16	0.16	0.12	0.16	0.14	0.15	0.14	0.17	0.17	20%	0.03	0	0.01	−0.03
mayfly larvae	0.12	0.11	0.15	0.09	0.09	0.11	0.09	0.12	0.11	0.11	0.12	0.14	17%	0.03	−0.12	0.01
newts	0.18	0.19	0.2	0.18	0.13	0.16	0.14	0.17	0.17	0.17	0.17	0.16	0.16	23%	−0.09	0.03
fish	0.14	0.13	0.16	0.13	0.11	0.14	0.12	0.13	0.14	0.14	0.14	0.15	0.14	0.16	14%	0.02
saucer bugs	0.17	0.17	0.2	0.15	0.11	0.16	0.12	0.11	0.15	0.12	0.16	0.16	0.13	0.17	0.15	25%

Fig. 3.C.2: Inter-species correlation coefficients from a) the community-only model and b) the full model. The upper-right triangle gives the mean of the posterior distribution over the correlation coefficients, with positive coefficients in green and negative coefficients in blue. The lower-left triangle gives the standard deviation of the posterior distribution over each coefficient. The diagonal gives the percentage of null deviance explained for each species.

Chapter 4

Methods for eliciting expert-opinion prevalence estimates and incorporating them in presence-only species distribution models.

Nick Golding & Bethan V. Purse

Authors' contributions: NG devised and implemented the methods and wrote the manuscript. BVP contributed to development of the methods and helped to revise the manuscript.

Abstract

Species distribution models (SDM) using occurrence and randomly selected background data are widely used in applied ecology. Predictions from these models are often interpreted as an estimate of the probability that the species is present. In fact, these predictions actually provide only a *relative* probability of presence or an index of habitat suitability. Knowledge of the species' prevalence is required to predict the absolute probability of presence. Since field data are usually not available from which to calculate this prevalence, it must be estimated from expert opinion.

We propose a framework of practical steps to overcome this problem, including: (i) a procedure for eliciting accurate estimates of prevalence from expert opinion; (ii) an approach to construct a valid probability distribution from these estimates which reflects their inherent uncertainty; (iii) a Bayesian approach to fitting presence-background SDMs which accounts for uncertainty in the prevalence estimate, applicable to any parametric SDM. We provide R code to implement these methods to enable their wider use.

Introduction

Species distribution models (SDMs) have become one of the most widely used analytical tools in ecology (Elith & Leathwick, 2009). They have been used to answer fundamental ecological questions (Sagarin *et al.*, 2006), though their greatest success has been in applied settings, such as conservation (Elith & Leathwick, 2006) and public health (Hay *et al.*, 2007). The most widely used SDMs attempt to describe a species' distribution as a function of its environment and are often referred to as niche models (Elith & Leathwick, 2009; Warren, 2012). Recent advances in SDM methodology have extended these approaches to incorporate intrinsic spatial effects (Vanhatalo *et al.*, 2012; Rangel, 2006), biotic interactions (Kissling *et al.*, 2011; Wisz *et al.*, 2013) and explicit biological processes (Kearney & Porter, 2009; Dormann *et al.*, 2012).

In some cases the data used to build these models are from surveys which record both presence and absence of the species. In these cases there are a number of different approaches which can be used to estimate the probability that a species is present a given location. Commonly though, the only data available at an appropriate spatial scale are records of sites where the species is known to have occurred (Graham *et al.*, 2004). Occurrence data are most often augmented with a random 'background' sample of the environmental conditions within the study area. We herein refer to this as 'presence-only' distribution modelling, whilst acknowledging that other methods exist which do not require a background sample (Walker & Cocks, 1991; Hirzel *et al.*, 2002). A widely used approach to presence-only modelling is the application of a standard presence-absence statistical model (Elith *et al.*, 2006), treating background data as absences. We term this the naïve approach after Ward (2007). Approaches such as MaxEnt (Phillips *et al.*, 2006), which explicitly consider background data, have also been widely used.

Often the aim of the SDM is to understand which environmental factors influence the species' distribution, or to produce maps predicting the relative suitability of habitat for the species. Where these are the sole aims of the study, both MaxEnt and

naïve models can be very successful, since they rank sites by their relative probability of presence (Elith *et al.*, 2006). In other cases, a map of the predicted *probability* that the species is present is required. Such maps may be used directly for policy making, or used as a part of further analysis, such as 'stacking' of distribution models to understand patterns of biodiversity (Benito *et al.*, 2013) or interactions between species (Broennimann *et al.*, 2012). In these cases the naïve approach performs poorly, since the prediction can have at best only a monotonic relationship with the true probability of presence (Phillips *et al.*, 2009). There is a common misconception that MaxEnt overcomes these problems - Yackulic *et al.* (2012) found that the majority of users interpreted predictions as probability of presence However MaxEnt actually generates predictions which are at best only proportional to the true probability of presence (Phillips *et al.*, 2006; Elith *et al.*, 2006). In many cases, the use distribution maps will be put to are unknown. It is therefore crucial to make clear precisely what they represent.

A number of approaches have been proposed to produce predictions of the species' probability of presence by incorporating the species' prevalence into the SDM (Phillips & Elith, 2013). Prevalence can be defined as the proportion of a set of randomly-allocated survey locations (in SDM this is commonly grid squares) in which the species is present. Prevalence and probability of presence can be difficult to define in presence-only SDMs, since they depend on the size of the grid squares, temporal scale and sampling approach (Elith *et al.*, 2011). Presence-absence data sampled randomly from the study area are rarely available (else presence-only methods would be unnecessary) so in practice prevalence estimates are likely to be elicited from experts with detailed knowledge of the species' ecology. These estimates are likely to be subect both to error and the experts' uncertainty. Methods for reducing error and incorporating uncertainty into model predictions would therefore be useful.

Here, we illustrate two forms of bias inherent in naïve models and propose a very simple approach which could be used to correct commonly used SDMs. We compare this simple correction with existing approaches at a range of prevalence estimates

and illustrate that it performs reasonably well, particularly at low prevalences. We propose methods to improve the accuracy of expert-opinion prevalence estimates and combine the uncertain estimates of multiple experts. Finally we demonstrate a flexible Bayesian statistical approach which can be used to fit any parametric SDM to presence-only data whilst incorporating uncertainty in the prevalence estimate into model predictions. Code is provided in the supplementary material to reproduce our simulations and figures using the statistical programming language R (R Development Core Team, 2012).

Bias in naïve models and MaxEnt

The presence-absence methods used in naïve models calibrate their predictions according the proportion of presence records in the data. As a result, the absolute value of predictions from naïve models are highly sensitive to the ratio of presence to background points used to fit the model (Ward, 2007). By contrast, MaxEnt explicitly considers background data and fits an exponential model which maximises the entropy between the presence and background samples. MaxEnt then rescales model predictions to fall between 0 and 1, using a rescaling parameter τ which reflects the species prevalence. By default MaxEnt assumes a prevalence of 0.5.

Fig. 4.2.1 illustrates bias in predictions from naïve models fitted using three different ratios of presence to background records and a MaxEnt model using the default prevalence of 0.5. Note that the naïve model with a ratio of 1:1 implicitly assumes a prevalence of 0.5, hence its similarity to the MaxEnt prediction. Because the MaxEnt model accounts for contamination of the background records with presence records (as discussed below) its estimate of the model slope is more accurate than that of the naïve model.

Whilst they produce biased predictions of the probability of presence, the naïve approach and standard Maxent approach are all capable of producing predictions of the rank-order of probability of presence. The performance of presence-only SDMs is commonly assessed using statistics such as the area under the reciever operating statistic curve (AUC, Fielding & Bell (1997)) and correlation coefficients (Elith *et al.*, 2006). These statistics measure only a model's ability to predict the rank order of probability of presence and essentially ignore the probability of presence suggested by the model (Lobo *et al.*, 2008). For example, the four predictive models in Fig. 4.2.1 had identical AUCs and Spearman's rank correlations of 0.78 and 0.36 for MaxEnt and naïve models respectively.

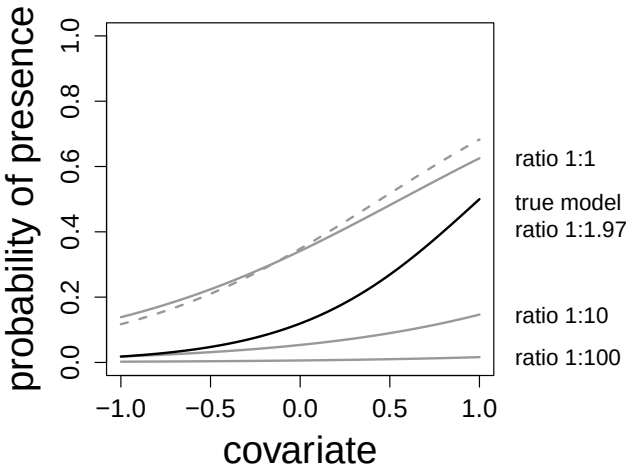


Fig. 4.2.1: Illustration of bias in naïve models and MaxEnt. Shown are naïve logistic regression models with three presence-background ratios (solid grey lines), a MaxEnt model fitted using only linear features, a ratio of 10 background records per presence and the default prevalence of 0.5 (dashed grey line) fitted to 1,000 presence and up to 100,000 background points simulated from a linear-logistic model with intercept -2 and slope 2 (solid black line, prevalence 0.34).

Calibration bias and contamination of controls in naïve models

Ward (2007) described the two sources of bias in naïve SDMs, which we refer to here as *calibration bias* and *contamination of controls*. Calibration bias arises because the ratio of occurrence to background samples is rarely the ratio of presence to absence records that would be expected in a random presence-absence sample. Contamination of controls is caused because background records are incorrectly assumed to represent sites where the species is absent. Because MaxEnt explicitly considers the presence-background case it is not susceptible to contamination of controls, though it is susceptible to calibration bias since it assumes a prevalence of 0.5 (Fig. 4.2.1).

The impact of both sources of bias on model predictions is dependent on the prevalence of the species. As the implicit prevalence of the model departs from the species' true prevalence, both naïve and MaxEnt models will consistently over or under predict the probability of presence. At higher prevalence a greater proportion of the randomly-selected background records will be undetected presences records,

whilst at low prevalence very few records will be undetected presence records and the model will be less biased

We ran a simulation to compare the relative importance of these two forms of bias on predictions from a naïve model under different species prevalences. Logistic linear regression models with one covariate were fitted to data generated from a true model with a slope of 2 and an intercept ranging from -2.9 to 2.9, resulting in prevalences from 0.08 to 0.92. The values of the covariate were sampled uniformly between -1 and 1. We fitted three types of model: a naïve model which was subject to both forms of bias, a model subject only to calibration bias and a model subect only to contamination of controls. The naïve model was fitted to 1,000 presence and either 10,000 or 100,000 background records (ratios of 1:10 and 1:100). The calibration bias model was fitted to 1,000 presence records and either 10,000 or 100,000 *absence* records. The contaminated controls model was fitted to 1,000 presence records and background records so that the ratio of presence to background records matched the expected ratio of presence to absence records in a random sample. We also fitted an unbiased model to 1000 presence points and the expected number of absence records for comparison. The predicted probability of presence from each of these models for a random sample of 50,000 records was compared to the true probability of presence by root mean squared error (RMSE).

Fig. 4.2.2 shows the results of this simulation. The calibration bias has a much greater impact on the predictive accuracy of the model than contamination of controls in our example. The relative importance of the two forms of bias on predictions from naïve models will depend both on the ratio of presence to background samples used and on the species' niche breadth. Note that in the calibration-corrected model prediction accuracies are higher at very high prevalences despite the background samples being heavily contaminated. This is because at very high and low prevalences most records are either presence or absence and the ability to correctly identify the probability of presence becomes more important than the ability to discriminate between presence and absences.

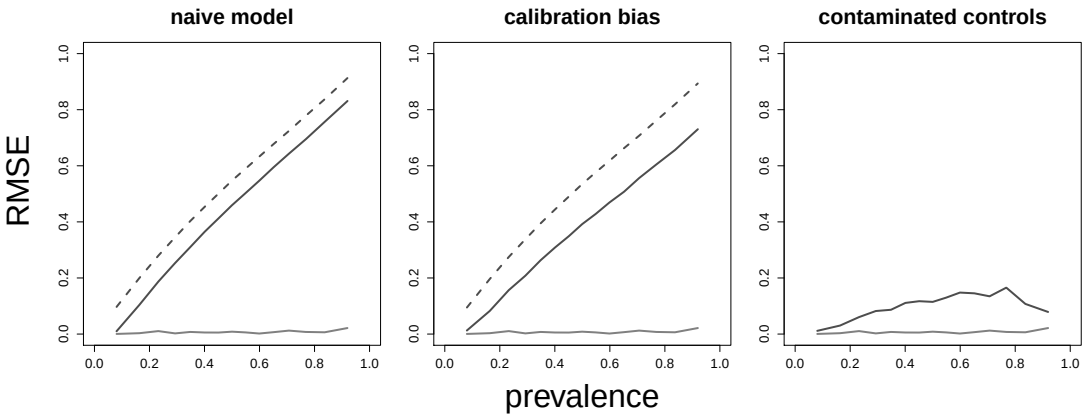


Fig. 4.2.2: Effect of calibration bias and contamination of controls on root mean squared error (RMSE) of predicted probability of presence. In the naïve and calibration bias plots, presence/background ratios of 1:10 and 1:100 are shown by solid and dashed lines respectively. The prediction error of an unbiased model is shown by a solid grey line.

Correction with a prevalence estimate

In order to produce predictions of probability of presence from presence-background data, methods are needed which account for both of the above forms of bias. Three methods have been proposed to do this using an estimate of the species' prevalence. They are: a likelihood function for logistic models (SC) proposed by Steinberg & Cardell (1992), an iterative expectation-maximisation algorithm (EM) proposed by Ward *et al.* (2009) and a scaled binomial loss function (SB) proposed by Phillips & Elith (2011). These three methods were reviewed by Phillips & Elith (2013). It is also possible to generate predictions of probability of presence from MaxEnt (ME) by replacing the default prevalence of 0.5 with a user-specified estimate of the species' prevalence. However, since ME fits an exponential model of habitat suitability then rescales this prediction to fall between 0 and 1, these are not robust maximum-likelihood predictions of probability of presence (Phillips & Elith, 2011).

A number of other studies have suggested that the species' prevalence estimate could be calculated from the data (Lancaster & Imbens, 1996; Lele & Keim, 2006; Divino *et al.*, 2011; Royle *et al.*, 2012). Whilst these approaches appear to perform very well on simulated datasets, they make very strong assumptions about model

structure (Ward *et al.*, 2009). They are highly susceptible to even minor violations of these assumptions, rendering them unsuitable for practical applications (Phillips & Elith, 2013). In practice prevalence estimates must be specified *a priori* and used to correct the model.

Calibration-corrected naïve model

As Fig. 4.2.2 illustrates, a large proportion of the prediction error in the naïve model is caused by calibration bias, rather than contamination of controls. We can therefore employ a very simple correction to this source of bias by using a subset of n_u background locations so that the ratio of presence to backround records is the same as would be expected in a random presence-absence sample. n_u is calculated as:

$$n_u = \frac{n_p(1-\pi)}{\pi} \tag{4.3.1}$$

where n_p is the number of prevalence records and π is the species prevalence.

We refer to this approach as the calibration-corrected naïve model (CC). This approach is similar to the case-control adjustment suggested by Ward *et al.* (2009) for logistic models. The CC model is very straightforward to implement for any statistical model designed for binary data (not just logistic models) and requires no programming on the part of the user. For SDM approaches which allow regression weights, this approach can be adapted to allow an arbitrary number of background records n_b by assigning each background record a weight of n_u/n_b, and presence records a weight of 1.

Whilst this correction to the standard naïve model allows the user make a valid prediction of the probability of presence of a species, it does not account for the 'contamination' of background samples with sites that are actually positive for the species. Although this contamination can lead to bias in the fitted model, this is only likely to cause an issue where the species prevalence is moderately high (since the rarer the species the fewer random background points, assumed to be absences, are likely to actually contain the species).

Comparison of calibration-corrected naïve model with other presence-only models

Using simulated data, we compared the performance of the CC model with other presence-only models. We investigate how their performance differs with species prevalence and when the prevalence estimate is incorrect. We fit the ME model (with only linear features and no regularisation) using the `dismo` R package (Hijmans *et al.*, 2012) and the SB, CC and naïve models using R's `glm` function (for SB using code given in Fig. 1 of Phillips & Elith (2011)). Since the EM and SC methods are more difficult to apply and implementations are not currently available, we provide R functions (`EM.linear` and `SC.linear`) for these in the supplementary material. These models were fitted to 100 presence points sampled from the same logistic linear models used in the simulation shown in Fig. 4.2.2 and 2,000 background points. Predictions from each model to a random sample of 50,000 records were compared with the true probability of presence by RMSE. We compared predictions of these models when provided with an inaccurate prevalence estimate, either 0.1 high or lower than the true prevalence (truncated between 0.01 and 0.99). These simulations were repreated 500 times and the mean RMSE from these simulations is shown in Fig. 4.3.1.

The presence-only models all had much higher predictive accuracy than the naïve model at all but very low prevalences. The comparative accuracy of the naïve model at very low prevalence arises because the prevalence implied by the 1:20 presence to background ratio (approximately 0.05) approaches the true prevalence. The predictive accuracy of all models was lower at high prevalences. In these cases, the species is likely to be present in most locations and the sample provides very little information with which to discriminate presence locations from the background. For very common species, samples of sites where the species is *not* found would be more useful than occurrence records, though this would obviously be difficult to achieve in practice.

Whilst the EM model performed well, particularly with prevalences above 0.5, though it was less accurate at prevalences below around 0.3. This effect was exacer-

bated when the prevalence was overspecified and ameliorated when it was underspecified. The reasons for this unexpected behaviour are unclear. The ME model was the least accurate of the presence-only models both when the prevalence was accurate and when it was underestimated (right hand panel Fig. 4.3.1). When the prevalence was overestimated (middle hand panel Fig. 4.3.1) ME models had *higher* predictive accuracy than when it was correctly specified, for prevalences less than 0.5. This is most likely a result of ME's rescaling of exponential predictions.

This rescaling procedure are likely the cause of the ME model's unusual behaviour. The SC model was overall one of the most accurate presence-only methods, though the average predictive accuracy dropped at high prevalences (>0.7, true prevalence) and particularly when the prevalence was overspecified. This was caused by the model occasionally specifying quite extreme parameter estimates and therefore predictions. The instability of this approach was noticed previously by Phillips & Elith (2013). In practice, application of the SC model should therefore be treated with caution. The SB model consistently made accurate predictions, and produced reasonable results even when the prevalence was mispecified. Whilst the predictive accuracy of the CC approach was lower than some of the other presence-only models the predictions were still correctly-calibarated (predicted probabilities were not consistently over- or underestimated). CC outperformed the ME model and its behaviour was consistent when the prevalence was mispecified.

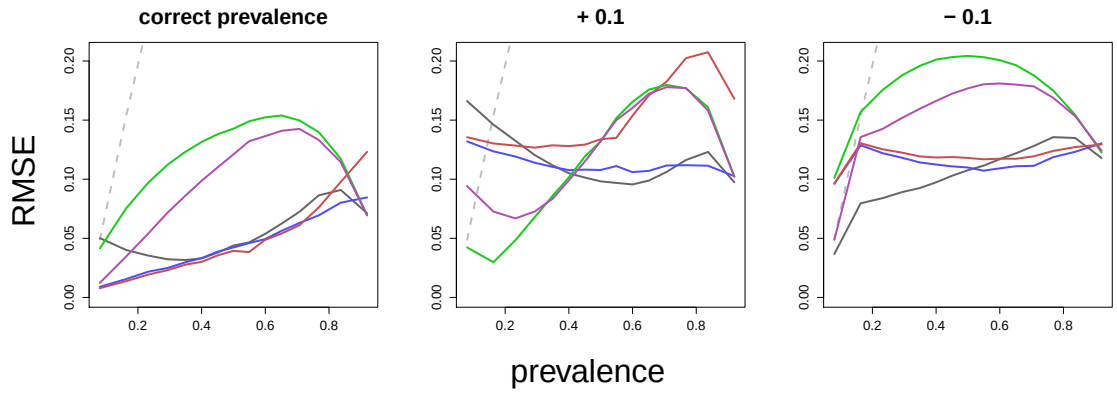


Fig. 4.3.1: Predictive accuracy of the expectation-maximisation (dark grey), MaxEnt (green), Steinberg & Cardell (red), scaled binomial (blue) and calibration-corrected naïve (purple) linear logistic models and a naïve model (dashed light grey).

Estimating prevalence from expert opinion

The previous section demonstrated the value of accurately specifying prevalence in species distribution models. Presence-only distribution modelling is typically applied when little or no presence-absence data is available. In a practical application of the methods discussed here, it is unlikely that suitable data therefore exist from which to estimate prevalence. Methods to elicit accurate estimates of species' prevalence from expert opinion would therefore be beneficial.

An obvious approach would be to simply ask a number of experts to provide estimates of the proportion of grid squares occupied by the species, given a clearly defined spatial extent, grid resolution and time period. To answer this question an expert may rely on their knowledge of the extent of the species' range in the study area, basing their estimate on the proportion of grid cells which fall within these limits. This approach could provide relatively accurate estimates if the resolution of the SDM were very low (large grid squares). At finer resolutions (small grid squares) this approach would likely lead to overestimates, since most species' distributions are relatively patchy within their range limits, and a smaller proportion of grid squares inhabited.

Incorporating additional knowledge of the species' ecology could help to improve this estimate. We propose to incorporate expert knowledge of the species' habitat preference by constructing a preliminary discrete-habitat based expert-opinion distribution model. Since prevalence can be calculated both as the proportion of locations inhabited and as the average probability of presence over these locations, averaging predictions from such model would provide an estimate of the species prevalence.

Whilst the continuous environmental covariates commonly used in SDM enable detailed predictions of the species' distribution, they may be difficult for the expert to interpret and relate to their knowledge of the species' ecology. Furthermore each covariate cannot be considered in isolation making the experts task yet more difficult. By contrast, discrete maps of habitat type are widely available and relatively straightforward to interpret in terms of a species' ecology. To construct a coarse expert-opinion distribution model, we can elicit from experts the proportion of grid squares (of a specified size and over a given time limit) of a given habitat type in which they would expect the species to be present. This information can be combined with the habitat map to produce a coarse map of the probability of presence of the species. The approach is ilustrated in Fig. 4.4.1.

In statistical terms, this approach is equivalent to considering prevalence as the *marginal* probability of presence, with the probability distribution of environmental conditions being marginalized out. Since we consider a mutually exclusive set of environmental conditions (habitat types) this marginalisation is calculated by a total probability sum:

$$\begin{aligned} p(y=1) &= \int p(y=1|x)\, p(x)\ \mathrm{d}x \\ &= \sum_{j=1}^{J} p(y=1|x_j^*)\, p(x_j^*) \end{aligned} \tag{4.4.1}$$

where x is a continuous measure of the environment and x^* is a discrete classification of the environment, with x_j^* being one of a set of J mutually exclusive discrete habitat types, $p(x)$ and $p(x^*)$ are the probability distributions over these two mea-

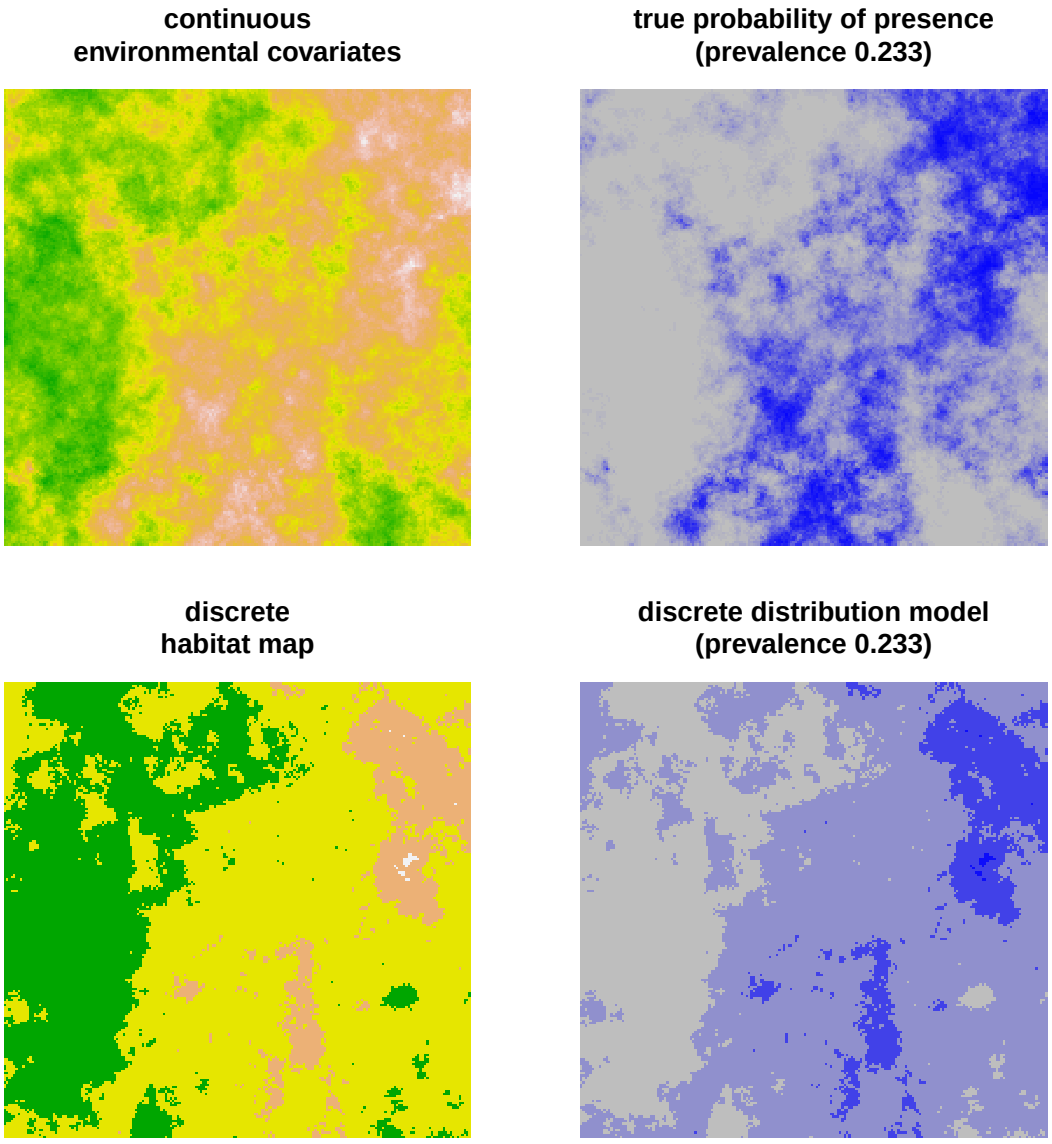


Fig. 4.4.1: Illustration of method for elicitation of species prevalence estimates from species experts. The species' true probability of presence across the study area (upper right) is assumed to depend on a set of continuous environmental covariates (upper left). We instead produce a discrete distribution model (lower right) for the species using a map of discrete habitat types (such as a land cover map, lower left) and expert-opinion estimates of the probability that the species would be present in a grid cell of a given habitat type.

sures of the environment (with $p(x_j^*)$ calculated as the proportion of the area covered by habitat type j), $p(y = 1|x)$ is the probability of presence conditional on continuous environmental variables (a species distribution model) and $p(y = 1|x_j^*)$ the probability of presence in grid square of habitat type j, which we elicit from experts.

Whilst it is likely that the habitat classes will be related to the continuous covariates used in the SDM (and those which drive the species' distribution) this is not required to estimate prevalence, provided the habitat-level probability estimates are accurate. The advantage of this approach is that instead of the entire prevalence estimate being subject to error in the expert opinion, only the habitat-specific prevalences are, with the proportional cover of habitat types being calculated from the habitat map. Whilst the habitat map may be subject to some error, this is likely to be far smaller than the error in expert opinion.

Simulation of prevalence estimates

To illustrate the advantage of this approach, we run three simple simulations, the results of which are shown in Fig. 4.4.2. For each run, we simulated species prevalences and proportional cover for a set of hypothetical habitat classes and combined these to derive a true (simulated) prevalence value. We simulated expert-opinion prevalence estimates by drawing values from a beta distribution estimated as in appendix 4.A, given a mode and 95% credible intervals (CIs) ± 0.3 (truncated to between 0.001 and 0.999). For the direct prevalence estimate we sampled estimates from a distribution with the true prevalence as its mode. For the discrete habitat method we sampled estimates for each habitat class with the class-level prevalence as the mode and summed across classes as in equation 4.4.1 to produce estimates of overall prevalence. We ran the simulation with theoretical habitat maps comprising 5, 10 and 20 habitat classes.

By breaking the expert opinion estimation down into discrete habitat types, the estimation error is minimized. The effect increases as the number of discrete habitat types increases. Note that this simulation assumes that error in the expert-opinion

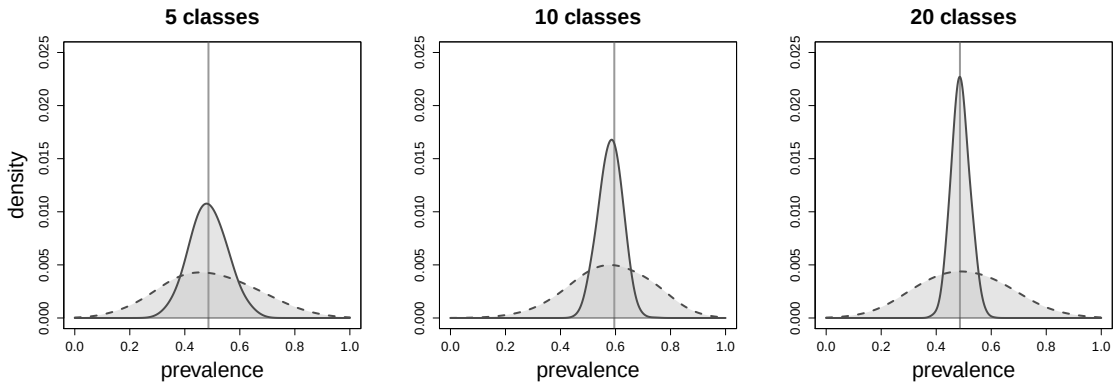


Fig. 4.4.2: Distribution of prevalence estimates using a direct estimate (dashed line) and the discrete habitat method described here (solid line) with different numbers of habitat classes. The density of estimates from 1000 simulations is shown and compared with the the true, simulated prevalence (vertical grey line).

estimates is the same whether for the overall prevalence or prevalence by habitat type. In practice, we might reasonably expect error to be lower for the habitat level prevalence estimates (particularly for those habitats which are known to be completely unsuitable), since more information is available from which to make an estimate.

Incorporating prevalence uncertainty in the model

Given a point-estimate of the species' prevalence, we can use any of the methods described above to build an unbiased SDM and generate predictions of the species' probability of presence. Such predictions are conditional on this single prevalence estimate and do not account for our uncertainty in it. It would therefore be sensible to incorporate this uncertainty into the model and generate predictions which are not conditional on a single estimate of the species prevalence.

Here we propose a Bayesian statistical approach to presence-only SDM which incorporates uncertainty in both the prevalence estimates and model parameters into model predictions. We outline a method to construct a valid prior probability distribution over the species prevalence parameter from uncertain expert-opinion estimates. We present a likelihood function which can be used to fit any parametric SDM to presence-background data. Finally, we demonstrate a Markov chain Monte Carlo procedure to fit and predict from such a model whilst accounting for uncertainty both in the prevalence estimate and model parameters.

Constructing a prior distribution

In order to fit a Bayesian statistical model we specify prior distributions over model parameters, which reflect our existing knowledge about the model. Whilst we can specify simple non-informative prior distributions over most model parameters, converting uncertain expert-opinion estimates into a valid probability distribution over the prevalence estimate is more difficult.

The beta distribution is a convenient model for the probability distribution over the prevalence estimate since it is well-understood and is bounded between 0 and 1. The shape of the distribution is controlled by two strictly positive parameters α and β. Unlike parameters of other commonly used distributions the relationship between the value of these parameters and the nature of the distribution are not easy to interpret. We are therefore unable to directly elicit estimates of these parameters

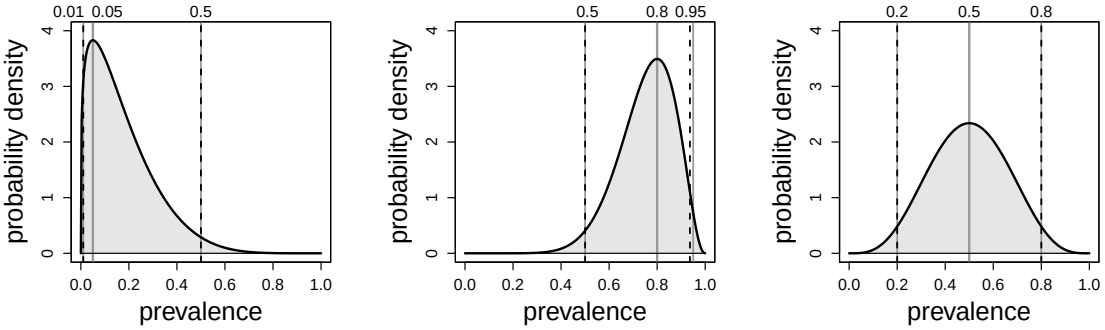


Fig. 4.5.1: Beta prior probability distributions for species prevalence fitted using three different expert-opinion modes and 95% credible intervals (grey lines and labels). The true 95% credible intervals of the distribution are indicated by dashed black lines. The estimated parameters for these distributions are (from left to right): $\alpha = 1.3$, $\beta = 6.0$; $\alpha = 9.9$, $\beta = 3.2$; $\alpha = 4.5$, $\beta = 4.5$.

from experts.

Instead, we elicit an estimate of the distribution's mode (the most probable prevalence value) and associated 95% CIs, reflecting the expert's uncertainty in this estimate. We then use these estimates to construct a valid beta distribution describing this uncertainty. Given these three estimated values, an exact solution to the two parametrs of the beta distribution is not guaranteed. We therefore use a numerical optimisation routine to find a valid beta distribution with the same mode and most similar 95% CIs (0.025 and 0.975 quantiles). This procedure is detailed in appendix 4.A and we provide an R function `betaPars` in the supplementary material to carry out this optimisation. Fig. 4.5.1 illustrates prevalence priors constructed using different modes and confidence intervals.

When eliciting expert-opinion prevalence estimates using a discrete distribution model as described above, we need to combine the uncertainty from a number of experts for several habitat types. We do so using a Monte Carlo integration procedure, outlined in appendix 4.B.

Adapting the presence-only likelihood function

Lancaster & Imbens (1996) proposed a likelihood function to fit unbiased binomial

models to presence-only data given a point-estimate of prevalence. This approach forms the basis for the expectation-maximisation algorithm of Ward (2007). An unfortunate limitation of this approach is that it is restricted to models which employ the logistic link function. We adapt this likelihood function to be suitable for any parametric model for binomial data by converting from the probability scale to the logistic scale before performing the presence-only correction. Our adapted likelihood function is:

$$\prod_{i=1}^{n} (p_i^*)^{z_i} (1 - p_i^*)^{1-z_i} \tag{4.5.1}$$

$$p_i^* = \frac{n_p}{n_p + \pi n_u} \frac{e^{\eta^*(x_i)}}{1 + e^{\eta^*(x_i)}} \tag{4.5.2}$$

$$\eta^*(x_i) = log\left(\frac{p_i}{1 - p_i}\right) + log\left(\frac{n_p + \pi n_u}{\pi n_u}\right) \tag{4.5.3}$$

$$p_i = f(x_i, \beta) \tag{4.5.4}$$

where z_i is a binary variable denoting whether observation i is a presence ($z = 1$) or background ($z = 0$) record, p_i^* is the probability that an observation represents a occurrence (rather than a background) record, p_i is the probability that the species is present and $f(x_i, \beta)$ is a model with vector of parameters β relating the covariates x_i to the probability that the species is present. Note that equation 4.5.2 is equivalent to equation 8 in (Ward *et al.*, 2009) and if the logistic link function is used, the entire likelihood function simplifies to equation 9 in the same article.

The formulation we give here can be used for a model with any link function suitable for binomial data, including the commonly used probit link or non-standard link functions such as the scaled Gaussian (McInerny *et al.*, 2011). This likelihood function can also be used to parameterise process-based models of species distributions which do not use an explicit link function (Dormann *et al.*, 2012). Because the likelihood function contains both convex and concave terms it can be difficult to

maximise using standard numerical approaches, particularly when the model has non-linear structure (Lancaster & Imbens, 1996; Ward, 2007).

Markov chain Monte Carlo procedure

Markov chain Monte Carlo (MCMC) has become a widely used approach for fitting Bayesian statistical models (Gelman *et al.*, 2004). The flexibility of MCMC enables the incorporation of various sources of uncertainty and for fitting non-standard models which are difficult to solve using conventional maximum-likelihood estimation techniques (such as the likelihood function we present above). We implement a Metropolis-Hastings sampler to fit a Bayesian logistic linear model to presence-background data using the adjusted likelihood function, whilst accounting for uncertainty in the prevalence estimate.

Unlike the other model parameters, the data contain no information on the value of the prevalence parameter π (subject to realistic assumptions). We therefore specify π *a priori* as in the other presence-only approaches discussed here. Unlike the other approaches, the stochastic nature of the MCMC algorithm allows us to consider different estimates of π within a single model. We do this by drawing a random sample of π from its prior distribution at each iteration of the algorithm. We use this estimate of π to sample the other model parameters according to the Hastings criterion. The resulting parameter estimates, and predictions generated from them account for all of the uncertainty in the prevalence estimate.

Unfortunately the non-standard approach we employ means that the model cannot be fitted using many of the popular software packages for MCMC, such as WinBUGS and OpenBUGS (Lunn *et al.*, 2009) We therefore provide an R function `MH` in the supplementary material which will perform the inference for a user-specified parametric model for binary data.

Using a simulated presence-background dataset we illustrate the MCMC sampler and the effect of incorporating prevalence uncertainty when the prevalence esimate is mispecified. We simulated 100 presence and 2,000 background points from the

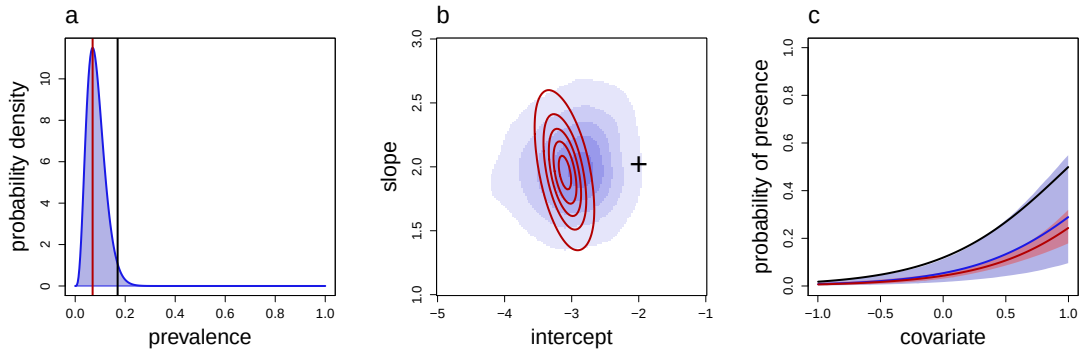


Fig. 4.5.2: Effect of prevalence uncertainty on model inference and predictions. **a)** Prior probability distribution over the species' prevalence (blue), with its mode (0.07) indicated by a red line and the true value (0.17) by a black line. **b)** Posterior density of parameter estimates from the model with prevalence uncertainty (blue shaded contours) and the model without (red contour lines). Contours show the 0.1 to 0.9 density regions, the true parameter values are indicated by a black cross. **c)** Posterior mean (solid coloured lines) and 95% credible intervals (shaded regions) of the predicted probability of presence from the model with prevalence uncertainty (blue) and without (red). The black line shows true probability of presence.

logistic linear model used to produce Fig 4.2.1. We used the MCMC sampler to fit a logistic linear model with a prevalence prior distribution with a mode of 0.07 and 95% CIs of 0.01 and 0.17 (beta distribution parameters: $\alpha = 4.8$, $\beta = 51.6$). The data had a true prevalence of 0.17, approximately at the upper 95% CI of the prior distribution. We fitted a second model which did not account for uncertainty in the prevalence estimate but assumed a fixed prevalence of 0.07. We ran both samplers for 50,000 iterations, after discarding 1,000 initial burn-in iterations which was sufficient to ensure that the Markov chains had converged and were well mixed.

The results of this simulation are shown in Fig. 4.5.2. Accounting for uncertainty in the prevalence estimate increased the uncertainty in parameter estimates. The model accounting only for uncertainty in model parameters produced predictive 95% CIs which were far from the true probability of presence. By contrast, the upper 95% CI of the model accounting for prevalence uncertainty incorporated the true probability of presence.

Discussion

In recent years, there has been an increasing awareness among SDM researchers that new methods are needed to produce robust distribution models from ubiquitous presence-only distribution data. As a result, a range of new approaches have been proposed (Phillips *et al.*, 2006; Ward *et al.*, 2009; Phillips & Elith, 2011; Divino *et al.*, 2011; Royle *et al.*, 2012) and existing methods from other fields of research rediscovered (Steinberg & Cardell, 1992; Lancaster & Imbens, 1996). Despite showing a great deal of potential to improve SDMs, these approaches have yet to become standard practice in applied distribution modelling. There are a number of potential barriers to the adoption of these methods, including a lack of software to fit robust presence-only models and the requirement to provide additional information in the form of a prevalence estimate.

The calibration-corrected naïve approach (CC) we introduce is immediately available for use with existing software - users need only select the number of background records to represent their prevalence estimate (using equation 4.3.1). Whilst it is subject to some bias, the CC approach generates predictions which are correctly calibrated and is capable of producing accurate predictions, particularly at low prevalences. Most of the methods we investigate here imply non-concave likelihood functions (Phillips & Elith, 2011) which can be difficult to maximise, leading to erratic predictions from some models. As the CC approach does not interfere with the model-fitting procedure it can be used to adapt any presence-absence model - regardless of the inference method.

The paucity of suitable data from which to estimate prevalence necessitates subjective estimates. Previous studies have investigated the potential to incorporate prior knowledge of species' responses to environmental variables into SDMs (Pearce *et al.*, 2001; Seoane *et al.*, 2005; Choy *et al.*, 2009; Murray *et al.*, 2009). However, we are not aware of any studies which have attempted to elicit expert-opinion estimates of species' prevalences. Based on our simple simulation, the discrete-habitat prevalence

elicitation method we propose here could drastically improve the accuracy of such estimates.

Given the subjectivity of expert-opinion estimates, a consideration of the uncertainty associated with them is crucial. Honest communication of the various sources of uncertainty in SDMs when generating predictions is crucial to allow proper assessment and interpretation of these models (Elith *et al.*, 2002). Bootstrapping and ensemble procedures have been used to quantify and incorporate multiple sources of uncertainty in SDMs (Elith & Leathwick, 2006). The Bayesian approach we propose incorporates uncertainty in both prevalence and parameter estimates into model predictions. This approach could be extended to incorporate other sources of uncertainty and prior ecological information.

The flexibility of MCMC for Bayesian inference means it can be used to parameterise complex species distribution models (Ovaskainen *et al.*, 2010; McInerny *et al.*, 2011). MCMC tends to be more computationally intensive than the numerical maximisation procedures currently used to fit SDMs. It also requires user input during the fitting process to ensure convergence of the sampler. This restricts wider use of this methodology to users with experience of using the technique and makes 'batch' fitting of models for multiple species impractical.

On the basis of the limited simulations we have performed, the methods we propose appear to be useful tools for generating robust predictions from presence-only SDMs. We have focussed on developing methods which can be easily applied using exisiting software and providing R code otherwise. We hope that this will enable further validation of these approaches and enable users of SDMs to apply them.

Acknowledgements

David Rogers provided helpful comments on the manuscript. We acknowledge funding from the NERC Centre for Ecology & Hydrology (Environmental Change Integrating Fund programme).

References

Benito, B.M., Cayuela, L. & Albuquerque, F.S. (2013) The impact of modelling choices in the predictive performance of richness maps derived from species-distribution models: guidelines to build better diversity models. *Methods in Ecology and Evolution*, **4**, 327–335.

Brent, R. (1973) *Algorithms for minimization without derivatives*. Dover books on mathematics. Dover Publications, Incorporated.

Broennimann, O., Fitzpatrick, M.C., Pearman, P.B., Petitpierre, B., Pellissier, L., Yoccoz, N.G., Thuiller, W., Fortin, M.J., Randin, C., Zimmermann, N.E., Graham, C.H. & Guisan, A. (2012) Measuring ecological niche overlap from occurrence and spatial environmental data. *Global Ecology and Biogeography*, **21**, 481–497.

Choy, S.L., O'Leary, R. & Mengersen, K. (2009) Elicitation by design in ecology: using expert opinion to inform priors for Bayesian statistical models. *Ecology*, **90**, 265–277.

Divino, F., Golini, N., Lasinio, G. & Penttinen, A. (2011) Data Augmentation Approach in Bayesian Modelling of Presence-only Data. *Spatial Statistics 2011: Mapping Global Change*.

Dormann, C.F., Schymanski, S.J., Cabral, J., Chuine, I., Graham, C.H., Hartig, F., Kearney, M., Morin, X., Römermann, C., Schröder, B. & Singer, A. (2012) Correlation and process in species distribution models: bridging a dichotomy. *Journal of Biogeography*, **39**, 2119–2131.

Elith, J., Burgman, M.a. & Regan, H.M. (2002) Mapping epistemic uncertainties and vague concepts in predictions of species distribution. *Ecological Modelling*, **157**, 313–329.

Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G.,

Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M.M., Townsend Peterson, A., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, Robert, E., Soberón, J., Williams, S., Wisz, M.S. & Zimmermann, N.E. (2006) Novel methods improve prediction of species distributions from occurrence data. *Ecography*, **29**, 129–151.

Elith, J. & Leathwick, J.R. (2006) Conservation prioritisation using species distribution modelling. A. Moilanen, K. Wilson & H. Possingham, eds., *Spatial Conservation Prioritization: Quantitative Methods and Computational Tools*, pp. 70–93. Oxford University Press, Oxford.

Elith, J. & Leathwick, J.R. (2009) Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 677–697.

Elith, J., Phillips, S. & Hastie, T. (2011) A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, **17**, 43–57.

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**, 38–49.

Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (2004) *Bayesian data analysis*. CRC press.

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A. (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution*, **19**, 497–503.

Hay, S., Graham, A. & Rogers, D. (2007) *Global Mapping of Infectious Diseases: Methods, Examples and Emerging Applications*. Advances in parasitology. Elsevier Science.

Hijmans, R.J., Phillips, S., Leathwick, J. & Elith, J. (2012) *dismo: Species distribution modeling*. R package version 0.7-17.

Hirzel, A., Hausser, J., Chessel, D. & Perrin, N. (2002) Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? *Ecology*, **83**, 2027–2036.

Kearney, M. & Porter, W. (2009) Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology Letters*, **12**, 334–50.

Kissling, W.D., Dormann, C.F., Groeneveld, J., Hickler, T., Kühn, I., McInerny, G.J., Montoya, J.M., Römermann, C., Schiffers, K., Schurr, F.M., Singer, A., Svenning, J.C., Zimmermann, N.E. & OHara, R.B. (2011) Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *Journal of Biogeography*, **39**, 2163–2178.

Lancaster, T. & Imbens, G. (1996) Case-control studies with contaminated controls. *Journal of Econometrics*, **71**, 145–160.

Lele, S. & Keim, J. (2006) Weighted distributions and estimation of resource selection probability functions. *Ecology*, **87**, 3021–3028.

Lobo, J.M., Jiménez-Valverde, A. & Real, R. (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145–151.

Lunn, D., Spiegelhalter, D.J., Thomas, A. & Best, N. (2009) The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*.

McInerny, G.J., Purves, D.W. & McIntyre, K.M. (2011) Fine-scale environmental variation in species distribution modelling : regression dilution , latent variables and neighbourly advice. *Methods in Ecology and Evolution*, **2**, 248–257.

Murray, J.V., Goldizen, A.W., OLeary, R.A., McAlpine, C.A., Possingham, H.P. & Choy, S.L. (2009) How useful is expert opinion for predicting the distribution of a species within and beyond the region of expertise? A case study using brush-tailed rock-wallabies Petrogale penicillata. *Journal of Applied Ecology*, **46**, 842–851.

Ovaskainen, O., Hottola, J. & Siitonen, J. (2010) Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology*, **91**, 2514–21.

Pearce, J.L., Cherry, K. & Whish, G. (2001) Incorporating expert opinion and fineâscale vegetation mapping into statistical models of faunal distribution. *Journal of Applied Ecology*, **38**, 412–424.

Phillips, S.J., Anderson, R.P. & Schapire, Robert, E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J.R. & Ferrier, S. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, **19**, 181–97.

Phillips, S.J. & Elith, J. (2011) Logistic methods for resource selection functions and presence-only species distribution models. *AAAI (Association for the Advancement of Artificial Intelligence*, pp. 1384–1389.

Phillips, S.J. & Elith, J. (2013) On Estimating Probability of Presence from Use-Availability or Presence-Background Data. *Ecology*.

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.

Rangel, T. (2006) Towards an integrated computational tool for spatial analysis in macroecology and biogeography. *Global Ecology and Biogeography*, **15**, 321–327.

Royle, J.A., Chandler, R.B., Yackulic, C.B. & Nichols, J.D. (2012) Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods in Ecology and Evolution*, **3**, 545–554.

Sagarin, R.D., Gaines, S.D. & Gaylord, B. (2006) Moving beyond assumptions to understand abundance distributions across the ranges of species. *Trends in Ecology & Evolution*, **21**, 524–30.

Seoane, J., Bustamente, J. & Diaz-Delgado, R. (2005) Effect of expert opinion on the predictive ability of environmental models of bird distribution. *Conservation Biology*, **19**, 512–522.

Steinberg, D. & Cardell, N.S. (1992) Estimating logistic regression models when the dependent variable has no variance. *Communications in Statistics - Theory and methods*, **21**, 423–450.

Vanhatalo, J., Veneranta, L. & Hudd, R. (2012) Species distribution modeling with Gaussian processes: A case study with the youngest stages of sea spawning whitefish (Coregonus lavaretus L. s.l.) larvae. *Ecological Modelling*, **228**, 49–58.

Walker, P. & Cocks, K. (1991) HABITAT: a procedure for modelling a disjoint environmental envelope for a plant or animal species. *Global Ecology and Biogeography*, **1**, 108–118.

Ward, G. (2007) *Statistics in ecological modeling; Presence-only data and Boosted MARS*. Ph.D. thesis, Stanford University.

Ward, G., Hastie, T., Barry, S., Elith, J. & Leathwick, J.R. (2009) Presence-only data and the em algorithm. *Biometrics*, **65**, 554–63.

Warren, D.L. (2012) In defense of 'niche modeling'. *Trends in Ecology & Evolution*, **27**, 497–500.

Wisz, M.S., Pottier, J., Kissling, W.D., Pellissier, L., Lenoir, J., Damgaard, C.F., Dormann, C.F., Forchhammer, M.C., Grytnes, J.A., Guisan, A., Heikkinen, R.K., Hø ye, T.T., Kühn, I., Luoto, M., Maiorano, L., Nilsson, M.C., Normand, S., Ockinger, E., Schmidt, N.M., Termansen, M., Timmermann, A., Wardle, D.A., Aastrup, P. & Svenning, J.C. (2013) The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biological Reviews of the Cambridge Philosophical Society*, **88**, 15–30.

Yackulic, C.B., Chandler, R.B., Zipkin, E.F., Royle, J.A., Nichols, J.D., Campbell Grant, E.H. & Veran, S. (2012) Presence-only modelling using MAXENT: when can we trust the inferences? *Methods in Ecology and Evolution*, **4**, 236–243.

Appendix 4.A Appendix A - Estimating the parameters of a beta distribution

Our aim is to estimate parameters α and β of a beta distribution with mode $\hat{\pi}$ and with 0.025 and 0.975 theoretical quantiles $\tilde{q}_{0.025,0.975}$ as similar as possible to the expert-opinion quantiles $q_{0.025,0.975}$. The mode of the beta distribution is given by:

$$\hat{\pi} = \frac{\alpha - 1}{\alpha + \beta - 2} \tag{4.A.1}$$

Given $\tilde{\alpha}$ (the estimate of α) and $\hat{\pi}$ we can solve for $\tilde{\beta}$:

$$\tilde{\beta} = \frac{\tilde{\alpha} - 1}{\hat{\pi} + 2 - \tilde{\alpha}} \tag{4.A.2}$$

We use a bounded line search algorithm (Brent minimization (Brent, 1973) using the R function `optimize`) to find a value of $\ln \tilde{\alpha}$ between -1 and 8 which minimises:

$$\sum |\tilde{q}_{0.025,0.975} - q_{0.025,0.975}| \tag{4.A.3}$$

The bounded search allows for stable maximisation and limits α to between 0.36 and 2981, sufficient to model a wide range of prevalence distributions.

Appendix 4.B Appendix B - Monte Carlo combination of prevalence estimates

Using the discrete-habitat prevalence elicitation method described here, we are provided with uncertain prevalence estimates for a number of habitat types from multiple experts. To incorporate this information into the model, we need to combine these estimates into a single prior probability distribution over the species' prevalence. We combine these multiple probability distributions using a Monte Carlo procedure, which we outline here.

For J habitat types and K experts, we have JK probability distributions $p(\pi_{j,k})$ over the species' prevalence. We approximate each of these as $p(\tilde{\pi}_{j,k})$ by constructing a beta distribution from the modes and 95% confidence intervals provided by the experts. For each expert k we:

- draw N random samples $\pi_{j\,k\,n}$ from each of the habitat-specific beta distributions $p(\tilde{\pi}_{j,k})$
- for each of these N sets of habitat- specific prevalence estimates we calculate the expected prevalence for the entire study area, as in equation 4.4.1:

$$\pi_{k\,n} = \sum_{j=1}^{J} \pi_{j\,k\,n}\; p(x_j^*) \tag{4.B.1}$$

  where $p(x_j^*)$ is the proportion of the study area covered by habitat type j.

We now have NK random samples $\pi_{k\,n}$ of π which we use to estimate $p(\pi)$. We estimate the mode of the distribution by kernel density estimation (using the `density` function in the base installation of R) and the 95% interval from the empirical 0.025 and 0.975 quantiles. Finally, we estimate the parameters α and β of a beta distribution approximation $p(\tilde{\pi})$ of the prevalence prior $p(\pi)$.

Chapter 5

GRaF: Fast and flexible Bayesian species distribution modelling using Gaussian random fields

Nick Golding & Bethan V. Purse

Authors' contributions: NG devised and implemented the model, designed and carried out the model comparison and wrote the manuscript. BVP contributed to the model comparison and helped to revise the manuscript.

Abstract

Species distribution models are very widely used in ecology and their predictions often inform policy decisions. It is therefore important to use methods with high predictive accuracy, that permit sources of bias to be taken into acount and enable biological interpretation.

We introduce a new approach for modelling species distributions using latent Gaussian random fields and provide an open source R package, GRaF, to allow ecologists to implement these models. We illustrate how GRaF works and describe some of its advantages over commonly used approaches, which include: i) flexible response terms, ii) the ability to account for imperfect occurrence records, iii) incorporation of prior knowledge of the species' ecology, iv) computational efficiency and v) automatic estimation of prediction uncertainty.

Using a dataset of 227 terrestrial vascular plant distributions within the UK we compare GRaF's predictive accuracy with that of widely used species distribution models and demonstrate that GRaF outperforms Boosted regression trees, Generalized additive models and MaxEnt. Finally, we discuss potential extensions for the approach, including spatially explicit and multi-species models and models for count data.

Introduction

Species distribution models (SDMs), in their basic form, attempt to model the distribution of species using environmental conditions as predictors. Typically these models make use of records of the distribution of the species in question and gridded datasets of environmental variables to generate maps of the species' predicted distribution. In recent years SDMs have become some of the most widely used methods in ecology (Elith & Leathwick, 2009), providing essential tools for both theoretical and applied research. Among other applications, SDMs are used to investigate drivers of global biodiversity patterns and to guide conservation policy and public health interventions (Lehmann *et al.*, 2002; Sinclair *et al.*, 2010; Sinka *et al.*, 2010).

A wide range of different approaches has been suggested for SDMs, ranging from relatively simple 'envelope' models and commonly used statistical methods such as logistic regression to more complex methods such as those developed in the field of machine learning (Elith *et al.*, 2006). These approaches have a number of features which determine their suitability to model species distributions including:

Predictive performance. Predictive accuracy is likely to depend on a number of factors, amongst which the ability to model complex (non-linear) interactions between drivers of species' distributions seems to be particularly important (Elith *et al.*, 2006). Preventing the model from overfitting to the training data (modelling random noise, rather than the true ecological response) will also increase predictive performance when the model is applied to new datasets (Wenger & Olden, 2012);

Imperfect data. SDMs are often applied to distribution records opportunistically collated from a variety of different sources, rather than from planned surveys. Such datasets are prone to various sources of error, such as observation bias, a lack of absence records and uncertainty in the location or reliability of individual records (Newbold, 2010; Elith *et al.*, 2010). Failure to account for

these sources of error can lead to biased model predictions.

Predicted distribution maps are often needed for species where few occurrence data are available (Pearson *et al.*, 2006). In these cases it may be useful to augment these limited data with existing knowledge of the species' ecology (Murray *et al.*, 2009).

Computational efficiency. SDMs are often required to be run in batch operations, where the distributions of multiple species are modelled at once. For routine analyses such as these, SDMs which are relatively quick to run and do not require much user input are advantageous.

Prediction uncertainty. SDM predictions typically represent the model's *best guess* at the species' distribution, given the occurrence and environmental data available. These predictions are usually subject to multiple sources of uncertainty in the data and model parameters. It is therefore desirable to provide maps quantifying uncertainty in these predictions (Elith *et al.*, 2002; Guisan & Zimmermann, 2000).

Gaussian random fields (GRFs; also known as Gaussian processes - particularly in the machine learning literature (Rasmussen & Williams, 2006)) provide a flexible approach to fitting complex statistical models. Most applications of GRFs in ecology require the use of computationally expensive Markov chain Monte Carlo samplers (Patil, 2007; Sigourney *et al.*, 2012). Whilst MCMC is a useful approach to fitting complex models it can be very time consuming and requires the user to supervise the model fitting process. These limitations make such procedures impractical for the routine fitting of SDMs (though see Vanhatalo *et al.* (2012)). We propose that latent GRF models fitted using efficient numerical appproximations can overcome these drawbacks and provide a solution to a number of issues inherent in species distribution modelling. A software package GRaF for the statistical programming language R (R Development Core Team, 2012) is provided to allow users to employ

these methods for fitting SDMs.

Below, we illustrate how GRF models work, demonstrate GRaF's solutions to some problems commonly encountered in SDM and compare GRaF's predictive ability with other commonly used approaches on a large dataset of known vascular plant occurrence records. The advantages and limitations of GRaF and potential avenues for future enhancements of the approach are discussed.

Gaussian random field models

While most statistical models attempt to describe the relationship between covariates and the response variable by parameterising some equation (e.g. linear regression), GRF models instead describe this relationship based on an assumption that observations with similar covariate values will yield a similar response value. The model then uses the available data to construct a normally distributed, correlated 'field' of variables which give rise to the observed response variable.

The approach of fitting models based on similarity (or dissimilarity) between sites, rather than fitting directly to the covariates, is shared by a number of related 'kernel methods' such as kernel regression (used in Generalised dissimilarity matrix models (Ferrier, 2002; Ferrier *et al.*, 2002)), kernel support vector machines (Evgeniou *et al.*, 2005) and kriging (Rogers & Sedda, 2012). The explicit Bayesian statistical treatment of GRF models enables a number of useful model-fitting procedures and extensions which would be difficult to implement for many of these other methods

The GRF approach is widely used in the field of model-based geostatistics (Diggle & Ribeiro Jr, 2007), where the covariates are the geographic location or the time of the observation and the GRF therefore models spatial or temporal correlation. GRaF takes this concept and applies it to *environmental* covariates to model the response variable - probability of presence.

A technical explanation of how this is achieved, and a statistical formulation of the GRaF model, are given in Appendix 5.A. Here we provide an illustration of how GRF models differ from other SDMs, using as an example the effect of temperature on the probability of presence of a hypothetical species.

Covariance function

In order to construct the GRF we first calculate the environmental distances between observations. In our example the environmental distances are simply the difference in temperature between each pair of sites (Fig. 5.2.1a). In a model with more co-

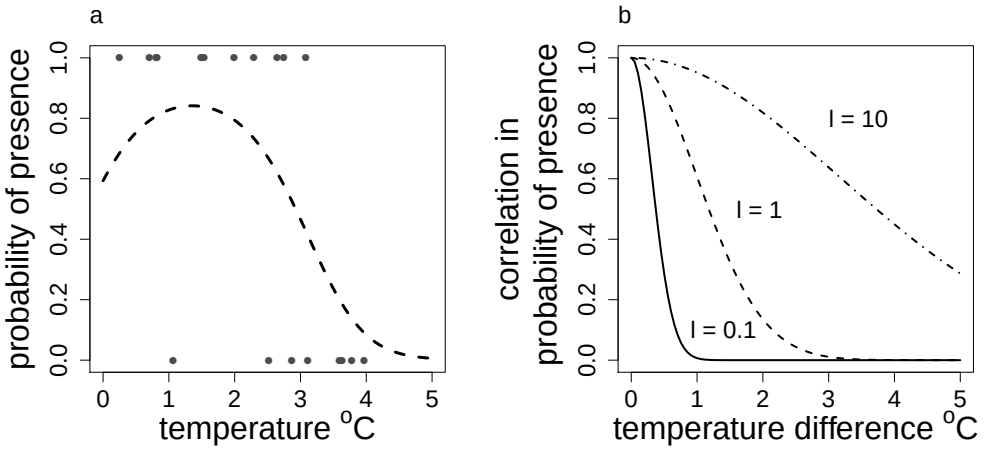


Fig. 5.2.1: Illustration of the covariance function. Panel **a** shows the observed presence-absence data (points) and the true underlying probability of presence given temperature (dashed line). Panel **b** shows how the covariance function relates differences in temperature between sites to correlation between probability of presence at these sites, given three different lengthscale parameters (discussed in the text). Models fitted to the data using these three lengthscales are shown in Fig. 5.2.2.

variates we would calculate the multi-dimensional Euclidean distance. We convert these distances into expected correlations in probability of presence between sites using a *covariance function*. There are a number of different covariance functions that we could use, but GRaF employs a squared-exponential covariance function since it is easy to parameterise and produces ecologically plausible smooth curves. As well as the distances between observations, we must supply the covariance function with a *lengthscale* parameter for each environmental covariate in the model. These lengthscales control how correlation between observations decays with environmental distance and therefore the complexity of fitted response curves. Fig. 5.2.1b shows how this function converts temperature difference to expected correlation given three different lengthscales. Assuming a lengthscale of one degree, the expected correlation between two observations with a one degree difference in tempreature is around 0.6, whereas with a difference of two degrees this drops to around 0.14. With a longer lengthscale, these expected correlations will be higher, resulting in a less complex fitted line (Fig. 5.2.2). In practice lengthscales do not need to be specified in advance

as they can be estimated from the data, though we may wish to inform the model of how likely different lengthscales are for the species being modelled.

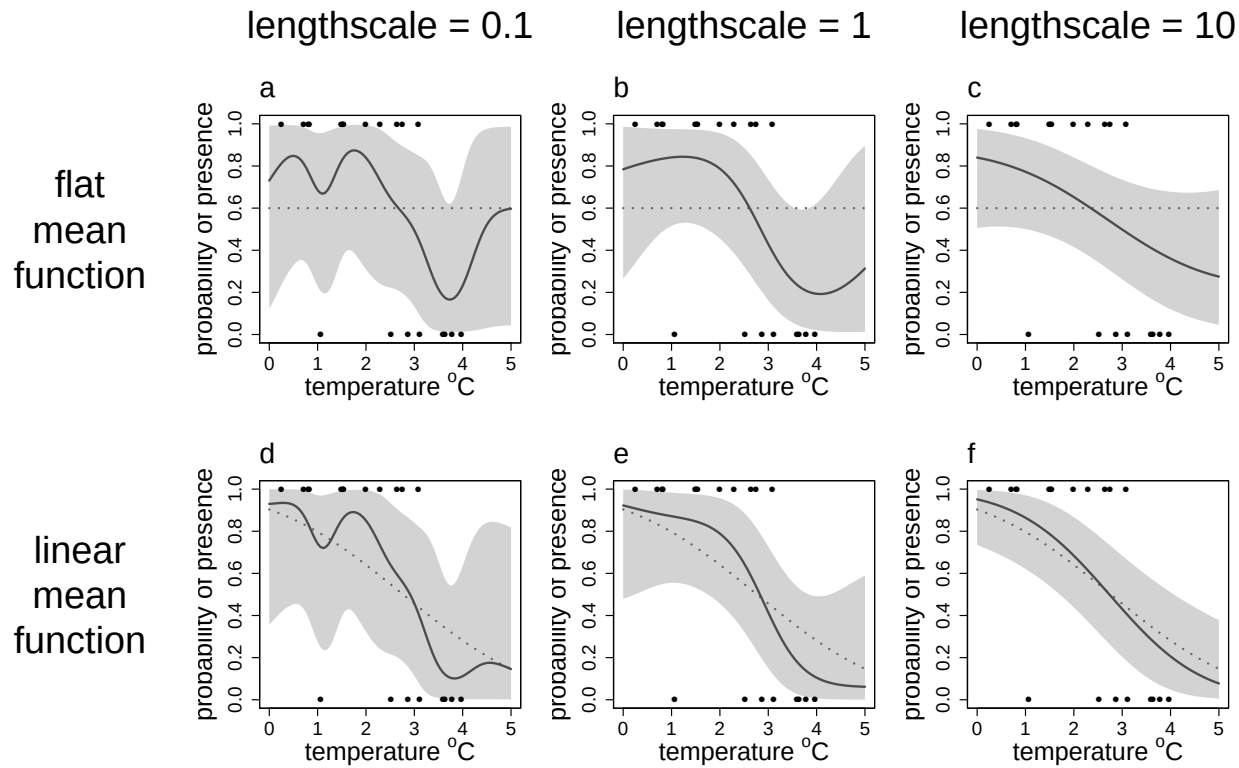


Fig. 5.2.2: The effect of the mean function and lengthscale on the fitted GRaF model. Shown are the observed data (points), the value of the mean function (dotted line), the probability of presence predicted by the GRaF model (solid line) and associated 95% credible intervals for this prediction (shaded grey area). Models are fitted with either the default flat mean function at the mean probability of presence (upper row) or a mean function representing some prior knowledge about how probability of presence relates to temperature, as described above (lower row).

Mean function

In addition to these expected correlations, we specify a *mean function*: an initial estimate of how the response variable changes with the covariates. If nothing is known about how the probability of presence of a species depends on temperature, a flat mean function is used, which assumes an equal probability of presence at all temperatures. If we have some prior knowledge that the species is more likely to be present at low temperatures than at high temperatures, we can incorporate this information into the model. For example, the mean function could be a linear model (with fixed parameters) relating temperature to probability of presence.

Fig. 5.2.2 demonstrates the effects of these two different mean functions on our model, with varying lengthscales. We can see from this illustration that where there are enough observations, the mean function has little effect on the fitted line, but where there are few datapoints, such as toward the limits of the recorded temperature range, the mean function determines the shape of the fitted response.

GRaF

Model structure

Machine learning algorithms such as boosted regression trees (BRT; Elith *et al.* (2008a)) have been shown to perform very well at predicting species distributions (Elith *et al.*, 2006). It seems likely that this performance is due to their ability to fit complex and highly non-linear responses to environmental covariates.

A drawback of BRT and similar methods is that they fit 'jerky' and biologically implausible predictive responses which may contribute to their tendency to overfitting to training data (i.e. they fit to noise in the data as well as the species' distribution, Wenger & Olden (2012)). By comparison, more traditional approaches such as univariate generalized additive models (GAM; Hastie & Tibshirani (1986)) fit more biologically realistic smooth functions. Whilst some implementations of GAM can fit multivariate smoothers (Wood, 2011), they perform poorly in more than a few dimensions so are unable to account for complex interactions.

GRaF offers an attractive solution to this trade-off between model flexibility and ecological realism by allowing for interactions between, and highly non-linear effects of, covariates, whilst fitting biologically plausible smooth predictive surfaces (see Fig. 5.3.1).

Uncertainty in occurrence data

Modellers often want to make high resolution predictions from high resolution gridded environmental data but are often hampered by the low resolution of the species occurrence data. This mismatch in resolution is problematic when the modeller needs to extract covariate values corresponding to each record, since there will be a number of different covariate values from which to choose. Thus the problem of uncertainty in the record location can be more usefully considered as uncertainty in the measured value of the environmental covariate for each occurrence record. A simple approach to this problem is to fit the model using the mean of the covariate values. Whilst

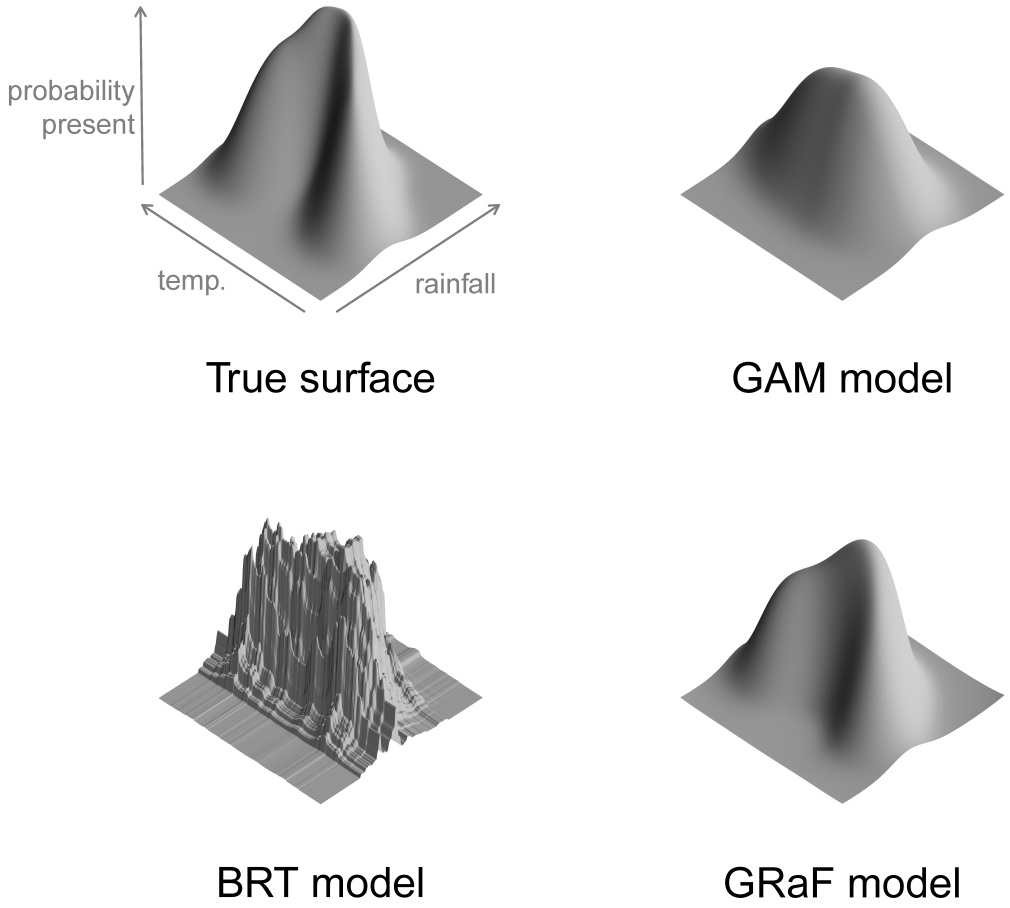


Fig. 5.3.1: Predictive surfaces fitted by boosted regression trees (BRT), a generalized additive model with univariate smoothers (GAM) and GRaF to simulated data with a strong non-linear interaction. The true surface represents the probability of presence of a hypothetical species in response to temperature and rainfall. Models were fitted to 1000 random presence/absence observations drawn from the true probability surface. R code used to fit these models is provided as supplementary material.

straightforward to implement, this approach ignores the uncertainty in the covariate and can lead to *regression dilution*, which dampens the apparent effect of a covariate on the species distribution.

The problem of regression dilution has been well studied in statistics (Frost & Thompson, 2000) and measurement error models have been proposed to deal with this bias in SDMs (McInerny *et al.*, 2011) and other ecological models (McNamara & Harding, 2004). As well as the mean of the environmental covariates for each record, measurement error models use an estimate of the error variance (which in the case of spatial error may be calculated from the multiple covariate values for each record) and use this information to correct for the regression dilution effect. Whereas most measurement error models are fitted by computationally expensive Markov chain Monte Carlo methods, GRaF fits these at negligible computational cost by accounting for this error within the covariance function (see Fig. 5.3.2). This approach is detailed in full in (Dallaire *et al.*, 2011).

GRaF also allows users to provide weightings to individual records in order to account for the variable reliability of different records or to account for observation bias (Phillips *et al.*, 2009) in a similar way to the bias grids available in other SDMs (Elith *et al.*, 2010).

Incorporating prior ecological knowledge

Often when modelling species distributions with few occurrence data there are other forms of information about the ecology of this or similar species which could be used to improve the model. For example experimental studies may have demonstrated a relationship between the species' ability to persist and some environmental gradient. In such cases it may be desirable to incorporate this *prior* knowledge of the species ecology into the model to augment the occurrence data. Unfortunately this is not easily accomplished in many current SDMs.

Bayesian statistical inference provides a convenient way of incorporating prior information of this sort into statistical models and has become increasingly popular

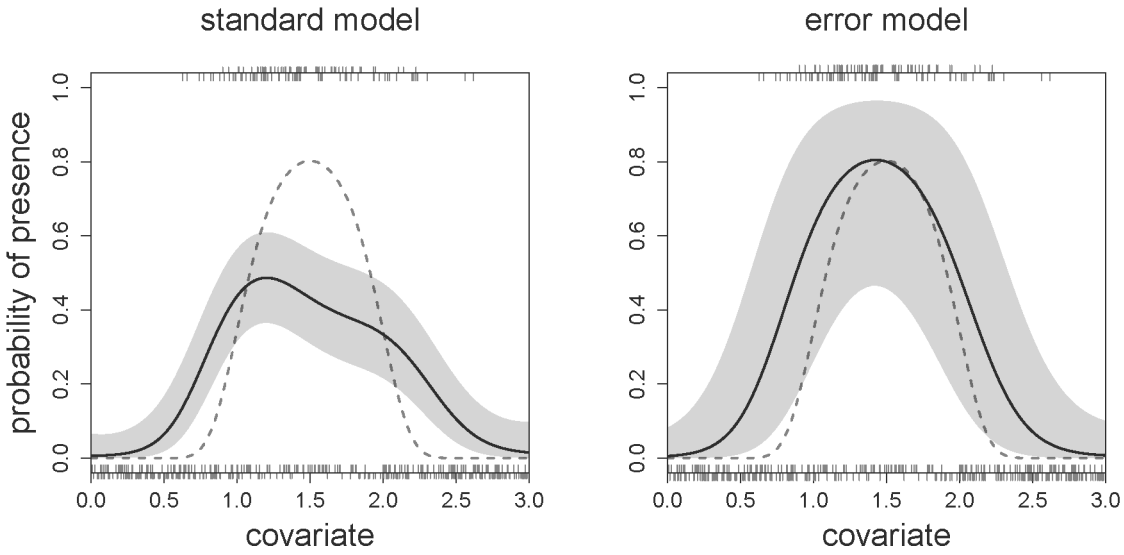


Fig. 5.3.2: Comparison of predictive surfaces fitted by GRaF models to simulated data with the covariate measured under error either ignoring measurement error (standard model) or accounting for it (error model). Solid lines give predictions and shaded regions represent 95% credible intervals. Three hundred presence-absence points were generated from the true model (dashed line) given the correct value of the covariate (tick marks outside box, presence on top line and absence on bottom line). Normally distributed random noise was then added to simulate covariates measured under error (tick marks insode box). Both models were then fitted to these data with measurement error. The error model was provided with the correct standard deviation of the error (0.5) whilst the standard model ignored the error.

in ecology (see e.g. McCarthy (2007)). In a Bayesian model the user specifies a prior probability distribution over each model parameter, representing their existing knowledge about what values of the parameter are likely. The model then compares this prior probability with the parameter estimate suggested by the data and produces a form of weighted average over the two; the posterior distribution. As the amount of data available to the model increases, the impact of the prior on the posterior diminishes.

GRaF allows the user to incorporate ecological knowledge into distribution models by manipulating two Bayesian priors: the mean function and the lengthscale hyperprior. The mean function acts as a prior over the whole SDM and can be used to incorporate specific knowledge of the species' response to environmental gradients, such as findings from experimental studies. The lengthscale hyperprior determines how likely different lengthscales are and can be used to inform the model how rapidly probability of presence is likely to change with different values of the environmental covariates.

If neither of these priors are manipulated by the user, GRaF fits a non-informative, flat mean function, as in Fig. 5.2.2 and an informative lengthscale hyperprior which represents ecologically plausible response curves (described in Appendix 5.A).

Uncertainty in model predictions

As with any model, predictions from SDMs are uncertain estimates of the probability of presence of the species. Where these predictions are to be used for some practical purpose it would be beneficial to provide maps representing the uncertainty in the predicted distribution map (Elith *et al.*, 2002). Such maps allow users to determine how much confidence they can place in a given prediction, information which is especially valuable where the predictons have policy implications.

SDM uncertainty estimates can be produced by bootstrapping data (Elith *et al.*, 2002) though this requires models to be run many hundreds of times and can therefore be computationally prohibitive.

GRaF models automatically produce estimates of uncertainty in model predictions, without the need for bootstrapping procedures. Fig. 5.3.3 illustrates predictions and associated uncertainty estimates from a GRaF distribution model of a plant species, Bog Myrtle (*Myrica gale*), in the UK.

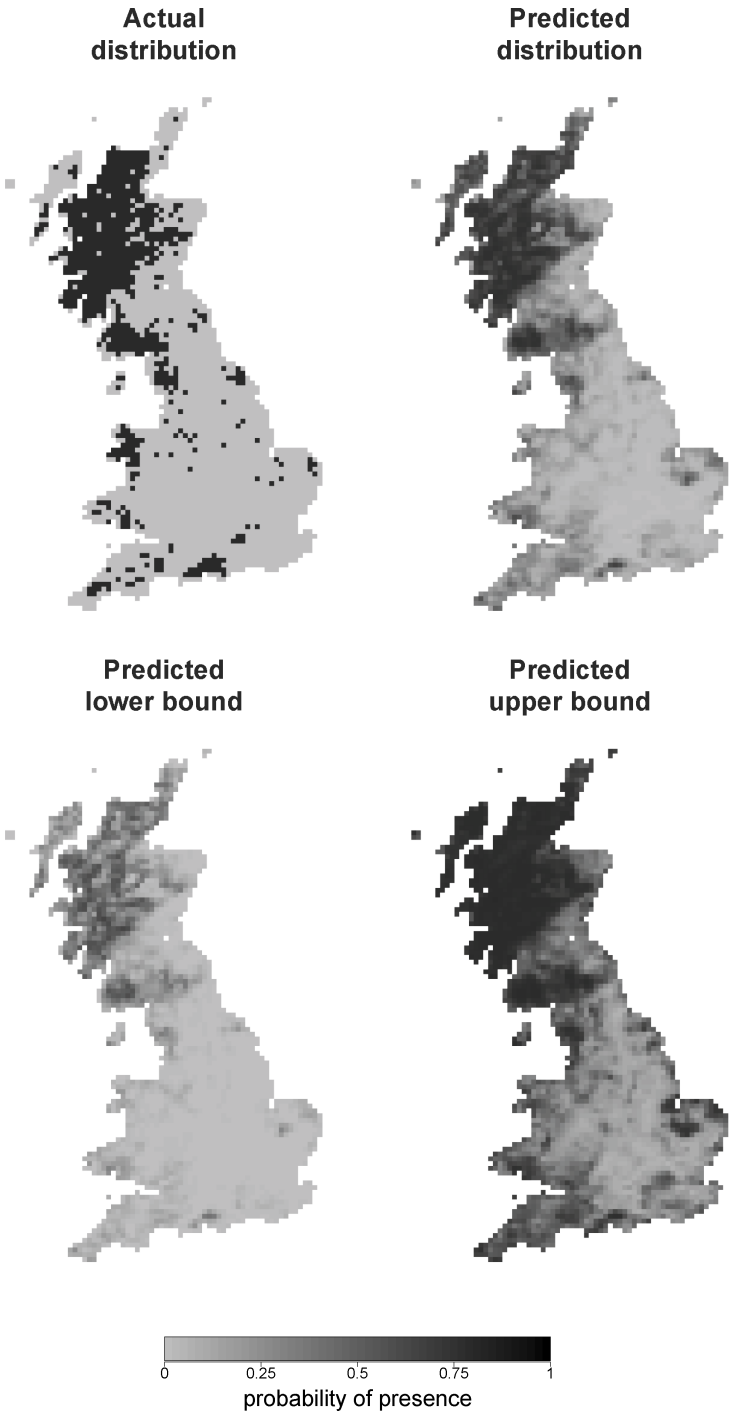


Fig. 5.3.3: True and predicted distributions of Bog myrtle (*Myrica gale*; prevalence 0.27) in Great Britain. Shown are the true distribution, the predicted distribution from a GRaF model (the Maximum *a posteriori* prediction), and lower and upper bounds on this prediction, representing our uncertainty in it (95% credible intervals, automatically generated by GRaF). The GRaF model was fitted to 300 presence-absence data points - around 10% of the dataset.

Comparison of GRaF with existing SDMs

We compared the predictive ability of GRaF with commonly used approaches for modelling species distributions from both presence-absence and presence-only data. All model fitting, prediction and statistical analysis were performed in R version 2.15. All R code and data used to carry out these analyses are provided as supplementary material.

Methods

Data

Gridded presence/absence maps of native terrestrial vascular plant species of Great Britain at 10 km resolution were obtained from the New Atlas of the British and Irish Flora (Preston *et al.*, 2002). Of the 1335 distributions in the original dataset, a subset of 227 species of different genera was selected for the model comparison. Criteria for selection of these species are outlined in Appendix 5.B. The distribution of British plant species is well characterized at this spatial scale, so records were assumed to represent known presence or absence which is essential to compare models fairly. The 227 plant species selected had a wide range of prevalences (proportion of grid squares occupied; ranging from 0.075 to 0.925, median 0.394), a factor known to influence the accuracy of SDMs (McPherson, 2004) and inhabited a wide range of habitats (see Fig. 5.B.1).

Maps of 10 indices of environmental conditions were used as covariates for model fitting and prediction. These were derived from a time series of satellite images of the UK by subsequent Fourier decomposition and principal components analysis to produce variables representing the major axes of environmental variation in the UK. The advantage of these abstract indices is that they compress a large amount of information on conditions and seasonality (the ten principal components explain 90% of variation in the Fourier variables) into relatively few orthogonal variables which enable us to make accurate predictions (Dormann *et al.*, 2008). Whilst they are

difficult to to interpret biologically, the first three indices appear to correspond to gradients from arable land to pasture, lowlands to uplands, and urban areas to rural areas. Details of this dataset and how it was produced are given in Appendix 5.B. All models were fitted using the full set of 10 covariates.

A total of 2774 10 km grid squares contained both distribution data and environmental data and were used to evaluate the different modelling approaches.

Presence/absence models

For the presence/absence comparison we compared GRaF with BRT and GAM. For each of the 227 species 300 grid squares (10.8% of available records) were selected at random and used to train each of the three models.

Model predictions for the remaining 2474 grid squares were then used to compare the predictive ability of the models. GRaF models were fitted using GRaF 0.1-0 (Golding, 2013), optimising the lengthscale parameters and otherwise using default settings. BRT models were fitted using the `gbm.step` function in dismo 0.7-17 (Hijmans *et al.*, 2012) with 5-fold cross validation, a tree complexity of 5, and an initial learning rate of 0.001. The learning rate was consecutively halved by an algorithm until a minimum of 1000 trees were fitted (in accordance with (Elith *et al.*, 2008a)). GAM models were fitted using gam 1.06.2 (Hastie, 2011) with univariate spline smoothers for each covariate and default settings.

Presence-only models

For the presence-only data we compared GRaF with MaxEnt (Phillips *et al.*, 2006), one of the most widely used approaches for modelling species distributions (Yackulic *et al.*, 2012).

In order to generate predictions about the probability of presence of a species using MaxEnt, the user must supply the model with an estimate of the species' prevalence in the study area. This information is not contained within the presence or background data and if not provided MaxEnt outputs can only be interpreted as a 'index of

habitat suitability' (Elith *et al.*, 2011). Equally, when fitting a model designed for presence/absence data to presence-background data it is important to select a ratio of background points to presence points equivalent to what what would be expected in a random sample of presence/absence data (Ward, 2007) in order to minimise bias in the predictions. In most cases it seems likely that the modeller could make an informed estimate of the prevalence of a species in the study area (though subject to some error) and correct the model, and this is the situation that we consider here.

For each species we selected 100 presence records at random from the dataset and 1500 randomly selected background points. We then fitted MaxEnt models to these 1600 data points and made predictions after correcting the model prevalence estimate using the species' prevalence in the entire presence-absence dataset. We fitted GRaF models to the same 100 presence records and a subset of the background points so that the ratio of presence to background points is equal to the prevalence. The remaining 1174 grid squares were then used to compare the models. GRaF models were fitted as for presence-absence data and MaxEnt models were fitted using dismo 0.7-17 with default settings.

Statistical analysis

For each model/species we compared the predicted probability of presence with true presence/absence on the withheld data. We used two metrics of predictive performance: the model deviance (Zuur *et al.*, 2007) and classification rate (proportion of grid squares correctly classified as presence or absence) using a threshold of 0.5. Since we produce predictions of the probability of presence, rather than a metric of suitability, a threshold of 0.5 is the natural choice, since it classifies the predictions according to the most probable case. We do not present results for the widely used AUC metric, since it is subject to a number flaws which limit its use for comparing between modelling approaches (Lobo *et al.*, 2008).

The results were analysed by mixed-effects regression, implemented using the nlme R package version 3.1-106 (Pinheiro *et al.*, 2012). In each regression the response

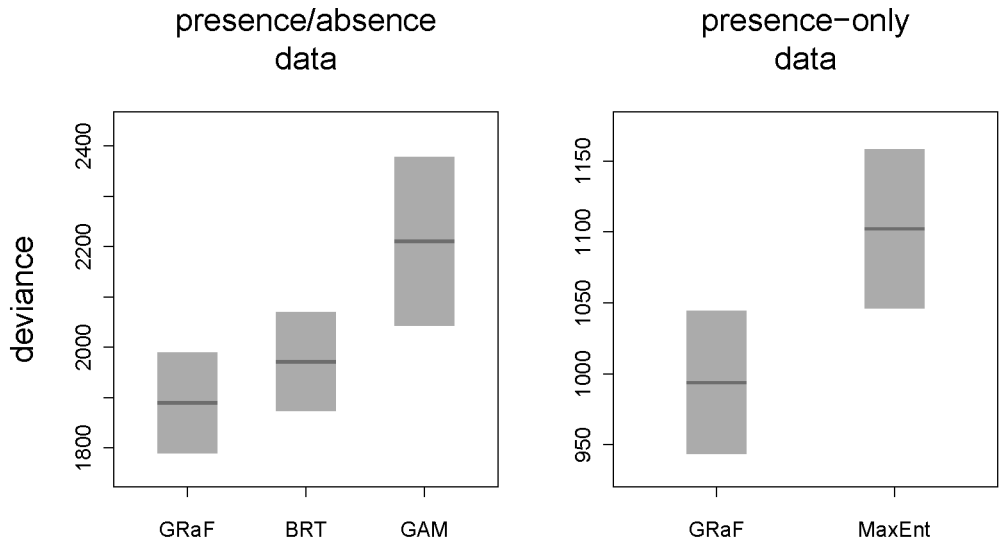


Fig. 5.4.1: Marginal deviance of model predictions to withheld training sets for presence/absence and presence-only data. Centre lines give means, boxes give standard deviations, lower deviance represents a better prediction. Presence/absence and presence-only models were tested on datasets of different sizes so deviances are not comparable between the two panels.

variable was the metric of predictive performance (deviance or classification rate) and the covariates were SDM (modelled as a fixed effect) and plant species (modelled as a random effect in order to account for the nested study design).

Marginal deviance scores were calculated from the residuals of null models with an intercept term and random effects terms for plant species, but no model term. These marginal deviances enable us to visualise the expected deviance from each SDM whilst removing species-level effects.

Results

GRaF models made more accurate predictions to the witheld data than other models for both presence-absence and presence-only data (Fig. 5.4.1). Predictive deviances for presence/absence GRaF models were 81.7 ($\pm$ 14.0 SE, $t_{452} = 5.83$, $p<0.0001$) lower than for BRT and 320.9 ($\pm$ 14.0 SE, $t_{452} = 22.90$, $p<0.0001$) lower than for GAM models. Predictive deviances for presence-only GRaF models were 108.5 ($\pm$

6.7 SE, $t_{226} = 16.13$, p<0.0001) lower than for MaxEnt models.

This pattern was mirrored in classification accuracies, with BRT classifying 0.6% (±0.06 SE, $t_{452} = 9.71$, p<0.0001), GAM 1.3% (±0.06 SE, $t_{452} = 20.40$, p<0.0001) and MaxEnt 3.8% (±0.33 SE, $t_{226} = 11.77$, p<0.0001) fewer grid squares correctly than GRaF's average rates of 82.5% and 79.5% respectively for the presence/absence and presence-only data.

GRaF models explained 28% of null deviance for presence/absence data and 20% for presence-only data.

Discussion

SDM comparison

In our comparison GRaF clearly outperformed a number of popular SDMs, including BRT which has been shown to be one of the best performing of existing SDM approaches (Elith *et al.*, 2008b).

In this comparison, we fitted each model following best-practice guidelines where available and default settings otherwise. We also compared models across a very large dataset of species distributions with varying prevalences. We therefore consider this to be a fair comparison of the SDMs considered. Further comparisons of these methods, with different datasets and modellers, would be useful in order to evaluate the performance of GRaF.

In the presence-only comparison, we provided GRaF and MaxEnt with information about the prevalence of the species. User specification of the species' prevalence is necessary in order to make a prediction about the species *probability of presence* (see Phillips & Elith (2013) and for a formal proof Ward (2007)). Unfortunately this important aspect of presence-only SDM appears to have been largely overlooked, with the output of uncorrected MaxEnt models in particular being very often misinterpreted as predictions of probability of presence (Yackulic *et al.*, 2012).

Advantages of a Bayesian approach

GRaF's ability to incorporate prior ecological knowledge and account for uncertainty in occurrence locations stem from its use of a Bayesian statistical approach. These features are likely to prove useful where there are few occurrence records (Murray *et al.*, 2009) and where there is a well-understood environmental driver of a species' distribution, such as a temperature limit on the distribution of a disease (Gething *et al.*, 2011). The same approach could be used to integrate process-based ecological models (Dormann *et al.*, 2012) with the more commonly used correlative SDMs.

By taking a Bayesian approach, GRaF can produce estimates of uncertainty in

model predictions, by considering a probability distribution over the predicted values (accounting for uncertainty in the shape of the GRF) rather than making a single 'best guess' prediction, as is the case with most SDMs.

Whilst GRaF predictions account for uncertainty in the shape of the GRF, they do not account for uncertainty in the lengthscale hyperparameters, but are conditional on an optimum estimate calculated from the dataset. Conditional posterior predictions are likely to satisfy most SDM users' requirements since most widely used SDMs do not provide any measure of prediction uncertainty, let alone accounting for uncertainty in hyperparameters. If required, predictions accounting for uncertainty in hyperparameters (or other variables, such as a prevalence estimate for presence-only data) can be approximated by numerical integration (e.g. a deterministic algorithm as in (Rue *et al.*, 2009) for models with parameters, or otherwise by Monte Carlo). Such a procedure will inevitably be more computationally intensive, but is likely to be far more efficient than alternative approaches such as MCMC.

Computational efficiency

GRaF models are fairly computationally efficient, with model fitting times to the 300 presence-absence datapoints in our comparison similar to those of BRT models (GRaF: mean 60.8s ± 21.41 SD cf. BRT: 43.7s ± 25.97s SD, n = 227). GRaF seems efficient for datasets of up to a few thousand datapoints and running our R implementation on a desktop computer is likely to be sufficiently fast for the majority of users.

Unfortunately, fitting times for GRaF models scale non-linearly with the size of the dataset (due to multiple matrix decompositions of $O(n^3)$ complexity), so for very large data sets, GRaF models can be disproportionately slow. For users who wish to model very large datasets efficiently, substantial speed-ups can be achieved by exploiting parallel computing. Since GRaF uses R's basic functions for linear algebra (which in turn call third-party linear algebra libraries) GRaF can be efficiently parallelised, without any additional coding, simply by linking R to a parallel linear algebra library

(see e.g. Schmidberger *et al.* (2009)).

Model complexity

GRaF can fit models with highly-complex interactions between covariates, a feature which probably contributes greatly to its predictive performance. When using complex models such as this it is important to avoid *overfitting* to the training data - finding patterns in random noise and therefore producing biased predictions to new datasets. Common approaches to dealing with this problem include stepwise covariate selection procedures, which seek to find a parsimonious subset of the available covariates which explain the data well; and regularization (e.g. the 'lasso' which is used in MaxEnt, among others (Tibshirani, 1996)) which is used during model fitting to include only the most important covariates.

GRaF tunes the model hyperparameters using the model marginal likelihood (Rasmussen & Williams, 2006) and therefore automatically reaches an optimal trade off between model fit and complexity. GRaF therefore reduces the influence of environmental covariates with little explanatory power and prevents model overfitting, without requiring the user to carry out a stepwise selection procedure.

A downside of fitting models with high-dimensional and non-linear interactions is that it becomes harder for the user to interpret the relationship between the species and its environment. This problem is not unique to GRaF, but is an inevitable trade-off when modelling complex data.

The GRaF R package provides functions to help users to interrogate GRaF models. The effects of individual covariates on probability of presence can be visualised, as can two-way interactions between covariates. Models can also be compared using deviance information criteria (Spiegelhalter *et al.* (2002), an equivalent of commonly used information criteria, such as Akaike's (Akaike, 1973), for hierarchical models) to quantify the relative importance of covariates in driving the species' distribution. A worked example modelling the distribution of Bog Myrtle and demonstrating these features is provided in appendix C.

Future work

Though GRaF is currently designed to fit the kinds of SDMs that are most commonly used at present, the approach could be extended in a number of ways. Since latent GRF models include, as a special case, a large subset of generalised linear mixed-effects models (Rue *et al.*, 2009) GRaF could easily be extended to model abundance data, nested study designs and spatio-temporal autocorrelation. Community GRaF models could also be implemented to allow users to fit models for whole communities of species, parameterising and accounting for correlations between their distributions (Wisz *et al.*, 2013; Kissling *et al.*, 2011). GRaF's open source R implementation will facilitate these extensions.

Acknowledgements

The authors acknowledge funding from the NERC Centre for Ecology & Hydrology (CEH) Environmental Change Integrating Fund Programme. We thank Chris Preston and David Roy from the Biological Records Centre, CEH, Wallingford and the Botanical Society of the British Isles for providing access to the plant atlas data and the maintainers of the CEH NEMESIS computing cluster. David Rogers, Miles Nunn, Luigi Sedda and David Harris provided helpful comments on the manuscript.

References

Akaike, H. (1973) Information Theory and an Extension of the Maximum Likelihood Principle. B.N. Petrov & F. Caski, eds., *Proceedings of the Second International Symposium on Information Theory*, pp. 267 – 281. Budapest.

Dallaire, P., Besse, C. & Chaib-draa, B. (2011) An approximate inference with Gaussian process to latent functions from uncertain data. *Neurocomputing*, **74**, 1945–1955.

Diggle, P.J. & Ribeiro Jr, P.J. (2007) *Model-based geostatistics*. Springer, New York.

Dormann, C.F., Purschke, O., García Márquez, J.R., Lautenbach, S. & Schröder, B. (2008) Components of uncertainty in species distribution analysis: a case study of the Great Grey Shrike. *Ecology*, **89**, 3371–86.

Dormann, C.F., Schymanski, S.J., Cabral, J., Chuine, I., Graham, C.H., Hartig, F., Kearney, M., Morin, X., Römermann, C., Schröder, B. & Singer, A. (2012) Correlation and process in species distribution models: bridging a dichotomy. *Journal of Biogeography*, **39**, 2119–2131.

Elith, J., Burgman, M.a. & Regan, H.M. (2002) Mapping epistemic uncertainties and vague concepts in predictions of species distribution. *Ecological Modelling*, **157**, 313–329.

Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M.M., Townsend Peterson, A., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, Robert, E., Soberón, J., Williams, S., Wisz, M.S. & Zimmermann, N.E. (2006) Novel methods improve prediction of species distributions from occurrence data. *Ecography*, **29**, 129–151.

Elith, J., Kearney, M. & Phillips, S.J. (2010) The art of modelling range-shifting species. *Methods in Ecology and Evolution*, **1**, 330–342.

Elith, J. & Leathwick, J.R. (2009) Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 677–697.

Elith, J., Leathwick, J.R., Hastie, T. & R. Leathwick, J. (2008a) A working guide to boosted regression trees. *Journal of Animal Ecology*, **77**, 802–13.

Elith, J., Leathwick, J.R., Hastie, T. & R. Leathwick, J. (2008b) Elith, Leathwick & Hastie A working guide to boosted regression trees - Online Appendices Page 1. *Journal of Animal Ecology*, **77**, 1–4.

Elith, J., Phillips, S. & Hastie, T. (2011) A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, **17**, 43–57.

Evgeniou, T., Micchelli, C. & Pontil, M. (2005) Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, **6**, 615–637.

Farr, T.G., Rosen, P.A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D. & Alsdorf, D. (2007) The shuttle radar topography mission. *Review of Geophysics*, **45**.

Ferrier, S. (2002) Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Systematic Biology*, **51**, 331–363.

Ferrier, S., Drielsma, M., Manion, G. & Watson, G. (2002) Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. II. Community-level modelling. *Biodiversity & Conservation*, **11**, 2309–2338.

Frost, C. & Thompson, S.G. (2000) Correcting for regression dilution bias: comparison of methods for a single predictor variable. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **163**, 173–189.

Fuller, R., Smith, G. & Sanderson, J. (2002) Countryside Survey 2000 Module 7. Land Cover Map 2000. Final Report. **2000**.

Gething, P.W., Van Boeckel, T.P., Smith, D.L., Guerra, C.a., Patil, A.P., Snow, R.W. & Hay, S.I. (2011) Modelling the global constraints of temperature on transmission of Plasmodium falciparum and P. vivax. *Parasites & Vectors*, **4**, 92.

Gilbert, M., Mitchell, A., Bourn, D., Mawdsley, J., Clifton-Hadley, R. & Wint, G.R.W. (2005) Cattle movements and bovine tuberculosis in Great Britain. *Nature*, **435**, 491–6.

Golding, N. (2013) *GRaF: Species distribution modelling using latent Gaussian random fields*. R package version 0.1-0.

Guisan, A. & Zimmermann, N.E. (2000) Predictive habitat distribution models in ecology. *Ecological modelling*, **135**, 147–186.

Hastie, T. (2011) *gam: Generalized Additive Models*. R package version 1.06.2.

Hastie, T. & Tibshirani, R. (1986) Generalized additive models. *Statistical Science*, **1**, 297–318.

Hijmans, R.J., Phillips, S., Leathwick, J. & Elith, J. (2012) *dismo: Species distribution modeling*. R package version 0.7-17.

Johnson, D., Hay, S.I. & Rogers, D.J. (1998) Contemporary environmental correlates of endemic bird areas derived from meteorological satellite sensors. *Philosophical Transactions of the Royal Society of London Series B, Biological sciences*, **265**, 951–959.

Kissling, W.D., Dormann, C.F., Groeneveld, J., Hickler, T., Kühn, I., McInerny, G.J., Montoya, J.M., Römermann, C., Schiffers, K., Schurr, F.M., Singer, A., Svenning, J.C., Zimmermann, N.E. & OHara, R.B. (2011) Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *Journal of Biogeography*, **39**, 2163–2178.

Lehmann, A., Leathwick, J.R. & Overton, J.M.M. (2002) Assessing New Zealand fern diversity from spatial. pp. 2217–2238.

Lobo, J.M., Jiménez-Valverde, A. & Real, R. (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145–151.

McCarthy, M. (2007) *Bayesian Methods for Ecology*. Cambridge University Press, Cambridge.

McInerny, G.J., Purves, D.W. & McIntyre, K.M. (2011) Fine-scale environmental variation in species distribution modelling : regression dilution , latent variables and neighbourly advice. *Methods in Ecology and Evolution*, **2**, 248–257.

McNamara, J.M. & Harding, K.C. (2004) Measurement error and estimates of population extinction risk. *Ecology Letters*, **7**, 16–20.

McPherson, J.M. (2004) The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology*, pp. 811–823.

Murray, J.V., Goldizen, A.W., OLeary, R.A., McAlpine, C.A., Possingham, H.P. & Choy, S.L. (2009) How useful is expert opinion for predicting the distribution of a species within and beyond the region of expertise? A case study using brush-tailed rock-wallabies Petrogale penicillata. *Journal of Applied Ecology*, **46**, 842–851.

Newbold, T. (2010) Applications and limitations of museum data for conservation

and ecology, with particular attention to species distribution models. *Progress in Physical Geography*, **34**, 3–22.

Patil, A. (2007) *Bayesian Nonparametrics for Inference of Ecological Dynamics*. Ph.D. thesis, University of California Santa Cruz.

Pearson, R.G., Raxworthy, C.J., Nakamura, M. & Townsend Peterson, A. (2006) Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *Journal of Biogeography*, **34**, 102–117.

Phillips, S.J., Anderson, R.P. & Schapire, Robert, E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J.R. & Ferrier, S. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, **19**, 181–97.

Phillips, S.J. & Elith, J. (2013) On Estimating Probability of Presence from Use-Availability or Presence-Background Data. *Ecology*.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core Team (2012) *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-106.

Preston, C., Pearman, D. & Dines, T. (2002) *New atlas of the British and Irish flora: an atlas of the vascular plants of Britain, Ireland, the Isle of Man and the Channel Islands*. Oxford University Press, Oxford.

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.

Rasmussen, C. & Williams, C. (2006) *Gaussian processes for machine learning*, volume 14.

Rogers, D.J., Randolph, S.E., Snow, R.W. & Hay, S.I. (2002) Satellite imagery in the study and forecast of malaria. *Nature*, **415**, 710–715.

Rogers, D.J. & Sedda, L. (2012) Statistical models for spatially explicit biological data. *Parasitology*, **139**, 1852–69.

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**, 319–392.

Scharlemann, J.P.W., Benz, D., Hay, S.I., Purse, B.V., Tatem, A.J., Wint, G.R.W. & Rogers, D.J. (2008) Global data for ecology and epidemiology: a novel algorithm for temporal Fourier processing MODIS data. *PloS ONE*, **3**, e1408.

Schmidberger, M., Tierney, L. & Mansmann, U. (2009) State of the Art in Parallel Computing with R. *Journal of Statistical Software*, **31**.

Sigourney, D.B., Munch, S.B. & Letcher, B.H. (2012) Combining a Bayesian nonparametric method with a hierarchical framework to estimate individual and temporal variation in growth. *Ecological Modelling*, **247**, 125–134.

Sinclair, S.J.S., White, M.M.D. & Newell, G.R. (2010) How useful are species distribution models for managing biodiversity under future climates. *Ecology and Society*, **15**.

Sinka, M.E., Bangs, M.J., Manguin, S., Coetzee, M., Mbogo, C.M., Hemingway, J., Patil, A.P., Temperley, W.H., Gething, P.W., Kabaria, C.W., Okara, R.M., Van Boeckel, T.P., Godfray, H.C.J., Harbach, R.E. & Hay, S.I. (2010) The dominant Anopheles vectors of human malaria in Africa, Europe and the Middle East: occurrence data, distribution maps and bionomic precis. *Parasites & Vectors*, **3**, 117.

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**, 583–639.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **58**, 267–288.

Vanhatalo, J., Veneranta, L. & Hudd, R. (2012) Species distribution modeling with Gaussian processes: A case study with the youngest stages of sea spawning whitefish (Coregonus lavaretus L. s.l.) larvae. *Ecological Modelling*, **228**, 49–58.

Ward, G. (2007) *Statistics in ecological modeling; Presence-only data and Boosted MARS*. Ph.D. thesis, Stanford University.

Wenger, S.J. & Olden, J.D. (2012) Assessing transferability of ecological models: an underappreciated aspect of statistical validation. *Methods in Ecology and Evolution*, **3**, 260–267.

Wisz, M.S., Pottier, J., Kissling, W.D., Pellissier, L., Lenoir, J., Damgaard, C.F., Dormann, C.F., Forchhammer, M.C., Grytnes, J.A., Guisan, A., Heikkinen, R.K., Hø ye, T.T., Kühn, I., Luoto, M., Maiorano, L., Nilsson, M.C., Normand, S., Ockinger, E., Schmidt, N.M., Termansen, M., Timmermann, A., Wardle, D.A., Aastrup, P. & Svenning, J.C. (2013) The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biological Reviews of the Cambridge Philosophical Society*, **88**, 15–30.

Wood, S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, **73**, 3–36.

Yackulic, C.B., Chandler, R.B., Zipkin, E.F., Royle, J.A., Nichols, J.D., Campbell Grant, E.H. & Veran, S. (2012) Presence-only modelling using MAXENT: when can we trust the inferences? *Methods in Ecology and Evolution*, **4**, 236–243.

Zuur, A., Ieno, E. & Smith, G. (2007) *Analysing ecological data*. Springer Verlag.

Appendix 5.A Statistical explanation and specification

5.A.0.1 Linear regression

The basic linear regression model can be written as follows:

$$\begin{aligned} \mathbf{y} &\sim \mathbf{N}(\mu, \Sigma) \\ \mu &= \alpha + \beta\mathbf{x} \\ \Sigma &= \mathbf{I}\sigma^2 \end{aligned} \tag{5.A.1}$$

where $\mathbf{y}$ is a vector of observed responses, which is assumed to be drawn from a multivariate normal distribution with mean vector μ and covariance matrix Σ. In order to describe how $\mathbf{y}$ responds to changes in the covariates $\mathbf{x}$, we model μ as a linear combination of $\mathbf{x}$ and vector of regression coefficients β with intercept α. We also assume that that each element of $\mathbf{y}$ is independent conditional on μ, so we can decompose Σ into an identity matrix $\mathbf{I}$ (having diagonal elements 1 and all other elements 0) and a single variance parameter σ^2.

5.A.1 Gaussian random fields

Gaussian random field models also assume that the observations $\mathbf{y}$ are drawn from a multivariate normal distribution, but rather than describing the relationship between $\mathbf{y}$ and $\mathbf{x}$ via the mean, μ is specified in advance and the relationship is modelled in terms of the *error* from the mean. Unlike in the linear regression model, these errors are assumed to be correlated, with the expected correlation between any two errors defined by a *covariance function*.

We can therefore specify a Gaussian random field model as:

$$
\begin{aligned}
\mathbf{y} &\sim \mathbf{N}(\mu, \Sigma) \\
\mu &= f(\mathbf{x}) \\
\Sigma &= g(\mathbf{x}, l) \\
ln(l_k) = \theta_k &\sim N(\mu_\theta, \sigma_\theta^2)
\end{aligned} \tag{5.A.2}
$$

where $f(\mathbf{x})$ is the mean function, evaluated at $\mathbf{x}$ and $g(\mathbf{x}, l)$ is a covariance function evaluated at $\mathbf{x}$, with lengthscale hyperparameter l. We place a normally-distributed hyperprior over each element of the natural log of l, which we denote θ, with mean μ_θ and standard deviation σ_θ. This hyperprior defines what values l can take, and therefore how smooth the fitted terms are, and is specified in advance, along with the mean function, to express the modeller's prior belief in the shape and flexibility of the GRF.

5.A.2 GRaF - specification

GRaF fits a *latent* GRF model, allowing us to consider binary (presence/absence) observations. This model is specified as follows:

$$
\begin{aligned}
\mathbf{y} &\sim Bernoulli(\mathbf{p}) \\
\mathbf{p} &= \Phi(\mathbf{z}) \\
\mathbf{z} &\sim \mathbf{N}(\mu, \Sigma) \\
\mu &= f(\mathbf{x}) \\
\Sigma &= g(\mathbf{x}, l) = e^{\sum_{k=1}^{m} -\frac{r_k^2}{2l_k}} \\
ln(l_k) = \theta_k &\sim N(\mu_\theta, \sigma_\theta^2)
\end{aligned} \tag{5.A.3}
$$

where $\Phi(.)$ denotes the cumulative density function of a standard normal distribution (the probit link function) which maps the latent GRF $\mathbf{z}$ to the vector of probabilities of presence $\mathbf{p}$ and $g(.)$ is the squared exponential covariance function

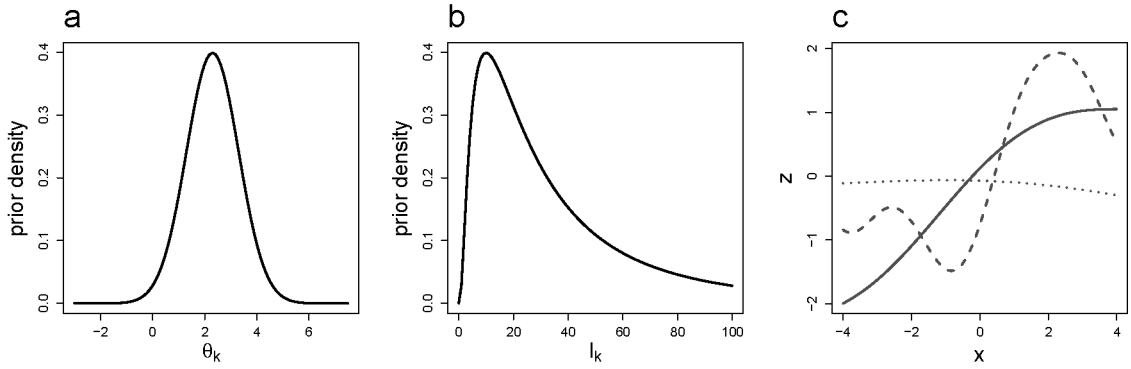


Fig. 5.A.1: Default hyperprior distribution ($\mu_\theta = ln(10)$, $\sigma_\theta = 1$) over each element of the hyperparameter θ. (**a**) Hyperprior distribution shown on the scale of θ_k. (**b**) Hyperprior distribution shown on the scale of l_k. (**c**) Examples of GRFs fitted using hyperparameters at the 0.025 ($l_k = 1.4$, dashed line), 0.5 ($l_k = 10$, solid line) and 0.975 ($l_k = 71.0$, dotted line) quantiles of the hyperprior.

for m covariates and r_k is the symmetric matrix of environmental distances between records for covariate k. When an error measurement is supplied with the covariates, the squared exponential covariance function of Dallaire *et al.* (2011) is used to incorporate this error.

GRaF users may specify a mean function $f(.)$ and parameters μ_θ and σ_θ of the hyperprior in order to reflect their prior knowledge about the species being studied. By default we set $f(\mathbf{x})$ to return the species' prevalence as calculated from the input data (a prior which assumes no effect of the covariates) and place an informative prior on θ with parameters $\mu_\theta = ln(10)$ and $\sigma_\theta = 1$ to represent our belief in relatively smooth effects of covariates on species' niches. This prior is illustrated in Fig. 5.A.1.

5.A.3 GRaF - fitting

Whilst there is a closed-form solution to the conditional posterior of a GRF with normally-distributed response data, finding a solution to the latent-Gaussian model is slightly involved. We use an iterative procedure to compute a Laplace approximation (Rasmussen & Williams, 2006) to the posterior distribution of $\mathbf{z}$, conditional on some value of the hyperparameters θ.

The posterior density of θ is given by:

$$p(\theta|y,x) \propto p(y|\theta,x)p(\theta) \tag{5.A.4}$$

where $p(\theta)$ is the prior probability of θ (defined by the hyperprior, described above) and $p(y|\theta,x)$ is the marginal likelihood of the model given θ and is calculated from the Laplace approximation. We use a standard numerical optimisation routine (BFGS, implemented using R's `optim` function) to find the peak of this density, giving us the maximum *a posteriori* estimate for θ which we use to approximate the conditional posterior distribution of the GRF and make predictions from the model.

Appendix 5.B Data for SDM comparison

5.B.1 Distribution data

Of the 1335 vascular plant distributions available in (Preston *et al.*, 2002), 1098 were selectively removed, leaving 227 distributions with which to compare SDMs. The rejection procedure was as follows:

1. Remove any distribution with a name containing the strings: 's.l.', 's.str', 'agg.', or 'sensu' since modelling of individual species, rather than species complexes, is more representative of standard applications of SDM. *95 removed*

2. Remove any species with more than 5% of records classed as 'non-native' or which were mapped as 'native' (where non-native occurrences were not considered and which may therefore contain non-native records; see Preston *et al.* (2002)), since modelling of non-equilibrium species distributions generally requires specific, case-by-case consideration (see e.g. Elith *et al.* (2010)). *251 removed*

3. Remove species for which distributions may be biased, according to expert opinion (Dr Chris Preston, *pers. comm.*). Reasons given were: incomplete recording, failure to distinguish between native and alien spp., species status debated or discredited. *7 removed*

4. Remove species classed as hydrophytes since their distributions are likely to be dependent on very local water conditions poorly described by remotely sensed imagery. *102 removed*

5. Retain one species at random from each genus to reduce the chance of phylogenetic correlation (and associated correlation of preferred habitat types) between species from reducing the independence of the data points. *522 removed*

6. Remove species with prevalence <0.075 or >0.925 to ensure that each distribution contained at least 200 presence and 200 absence records, enough to enable

both training and testing of models. *131 removed*

The distributions of the remaining 227 vascular plant species showed no clear habitat bias, though species richness displayed a slight north-south gradient (Fig. 5.B.1).

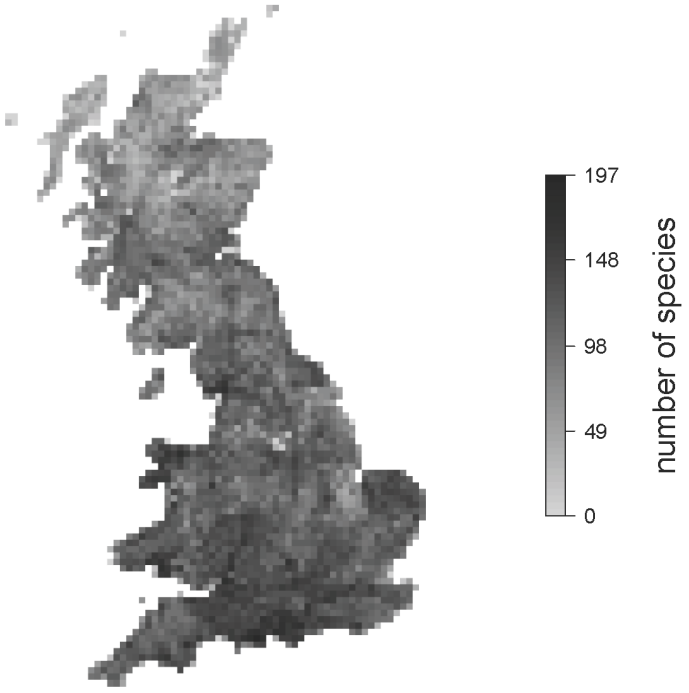


Fig. 5.B.1: Density of presence records for the 227 vascular plant species used to compare SDMs

5.B.2 Environmental covariates

In order to build distribution models for the plant data, we require gridded maps of environmental data. We produced maps of the major axes of enviromental variation in Great Britain, derived from a 7-year time series of high-resolution satellite images.

Unlike maps of land cover or meteorological conditions from weather stations which are commonly used for building SDMs, satellite imagery has the advantages that it is directly recorded at high resolution (rather than being interpolated from a small number of sites) and represents a continuous measure of environmental conditions (rather than subjective land cover classes). Whilst a single satellite image

provides useful information about environmental conditions, by using a time series of images, we are able to capture additional detailed information of seasonal change.

High resolution (approx. $200m^2$) satellite images of Great Britain, recorded by the moderate-resolution infrared spectrometry (MODIS) sensor on NASA's Terra and Aqua satellites every 8 days between 2001 and 2007 were obtained. For each pass, we obtained images of the following three bands: mid infra-red (MIR; raw output of wavelengths 3-5 μm); enhanced vegetation index (EVI; a metric of vegetation cover) and normalized vegetation index (NDVI; a metric of vegetation cover complementary to EVI). In order to render this huge dataset (containing more than 800 images, each with around 15 million pixels) usable for SDM, it was necessary to compress this information into a smaller set of gridded maps. The imagery were first decomposed using a temporal Fourier analysis (TFA) procedure, designed specifically for MODIS data and described in detail in (Scharlemann *et al.*, 2008). This procedure returns, for each pixel and each band, 10 variables: the mean, variance, maximum and minimum of the band and 6 Fourier components: the phase and amplitude of annual, bi-annual and tri-annual cycles. These TFA variables together describe the temporal pattern of each band within each pixel, and have been successfully used in SDMs (Rogers *et al.*, 2002; Gilbert *et al.*, 2005; Johnson *et al.*, 1998). Whilst the Fourier components within each band are orthogonal to one another, correlations remain between bands and with the other TFA variables.

We use principle components analysis (PCA) on these 30 TFA variables, plus an elevation map at the same resolution from the Shuttle Radar Topography Mission (Farr *et al.*, 2007), to derive a set of 31 orthogonal principle components which describe the *major axes* of environmental variation in the TFA variable dataset. The first 10 principle components represent around 90% of the total variation in the TFA dataset (see Table 5.B.1) and we retain these for use in the SDM comparison. In order to model the plant distribution data at 10km resolution, we resampled the maps of these 10 components from 200m resolution to 10km resolution (the spatial distribution of these principal components is shown in Fig. 5.B.3).

Whilst it would have been possible to carry out the PCA on the original MODIS time series, this would have required the highly computationally-intensive decomposition of a matrix of around 12 billion elements. Carrying out the PCA on the TFA variables (which had already been produced prior to this study) required decomposition of a much smaller matrix.

Despite the comparatively high information content of these principle components, they have the disadvantage that they are not as directly ecologically interpretable as other commonly used environmental variables, such as land cover classes. In order to interpret the environmental gradients that these indices represent, we compare each of the 10 principal components with broad land cover classes at the same resolution obtained from the CEH land cover map 2000 (Fuller *et al.*, 2002) (Table 5.B.1). The first three components appear to represent gradients from Arable land to pasture, from lowlands to uplands and from urban areas to coniferous woodlands respectively.

The ability of these components to describe environmental variation becomes more clear when they are combined. Fig. 5.B.2 combines the first three principal components. It is clear in this figure that arable land (red), pasture (yellow-green), uplands (purple) and urban areas (turquoise) are well described by the first three principal components. The additional components will combine to describe finer details of the environment which may be equally as important for describing species' niches.

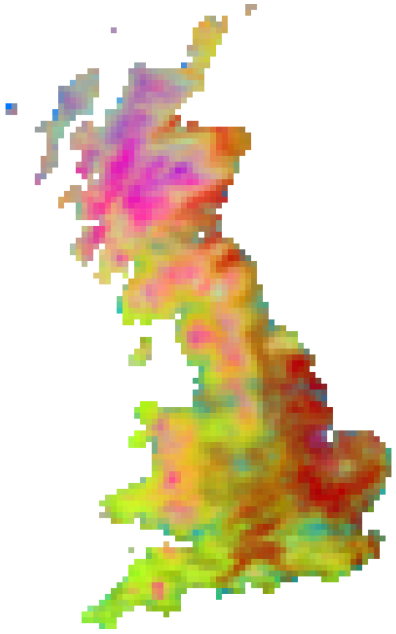


Fig. 5.B.2: Composite image showing the spatial distribution of principal components 1 (green), 2 (blue) & 3 (red) in Great Britain. These three components account for 60% of the total variance in the TFA dataset.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Broadleaved/mixed woodland	0.10	-0.52	-0.03	0.13	0.21	-0.12	0.10	0.04	0.21	0.37
Coniferous woodland	0.08	0.19	**0.59**	0.13	-0.09	-0.02	0.10	-0.22	**0.26**	0.21
Arable & horticulture	**-0.29**	**-0.86**	-0.41	-0.11	0.10	**-0.29**	-0.05	0.04	-0.03	0.07
Improved grassland	**0.42**	-0.56	0.11	0.22	0.22	-0.17	**0.21**	0.11	0.25	0.24
Semi-natural grassland	-0.06	0.14	0.49	**0.36**	0.05	**0.20**	0.18	-0.09	0.11	**0.40**
Mountain, heath & bog	-0.04	**0.64**	0.58	0.22	**-0.24**	0.12	0.10	**-0.24**	0.07	0.12
Built-up areas & gardens	0.04	-0.62	**-0.50**	-0.02	**0.34**	-0.11	0.08	0.15	-0.06	0.22
Standing open water	-0.15	0.26	0.10	0.06	-0.10	0.20	-0.06	-0.06	-0.10	0.11
Coastal	0.19	0.22	-0.24	-0.25	0.12	0.08	-0.14	0.17	**-0.29**	-0.49
Oceanic seas	0.19	0.36	-0.18	**-0.31**	0.05	0.12	**-0.18**	**0.18**	-0.27	**-0.52**
Proportion of variance explained	0.27	0.22	0.11	0.08	0.05	0.04	0.03	0.03	0.02	0.02
Cumulative proportion explained	0.27	0.49	0.60	0.68	0.73	0.77	0.80	0.83	0.85	0.87

Table 5.B.1: Spearman's rank correlation coefficients between principal components and broad land cover classes from LCM 2000, and proportion of variance of TFA components explained by each component. The strongest positive and negative correlation coefficients between each principal component and land cover classes are given in boldface.

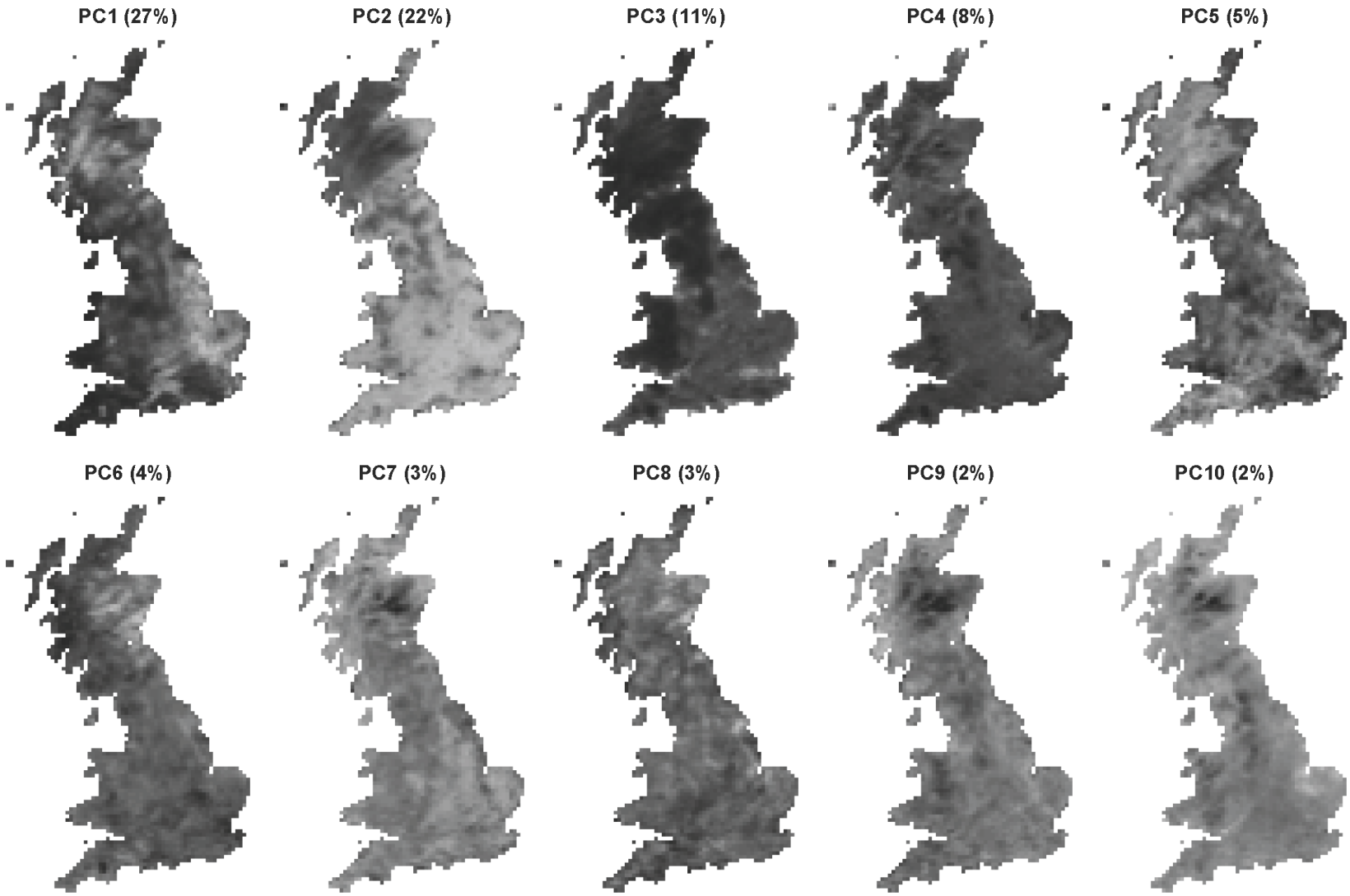


Fig. 5.B.3: Spatial distribution of principal components in Great Britain. Darker shades represent higher values of the component. Proportion of variance in TFA dataset explained by each component is given in brackets.

Appendix 5.C Demonstration of GRaF R package

Introduction

We demonstrate here some features of the GRaF R package for fitting species distribution models using Gaussian random fields. As an example, we model the distribution of Bog Myrtle (*Myrica gale*) in Great Britain using data detailed in the methods section and Appendix B. We assume that both the GRaF package and the raster package are installed and that the file `BogMyrtleData.RData` is the working directory.

Loading and extracting data

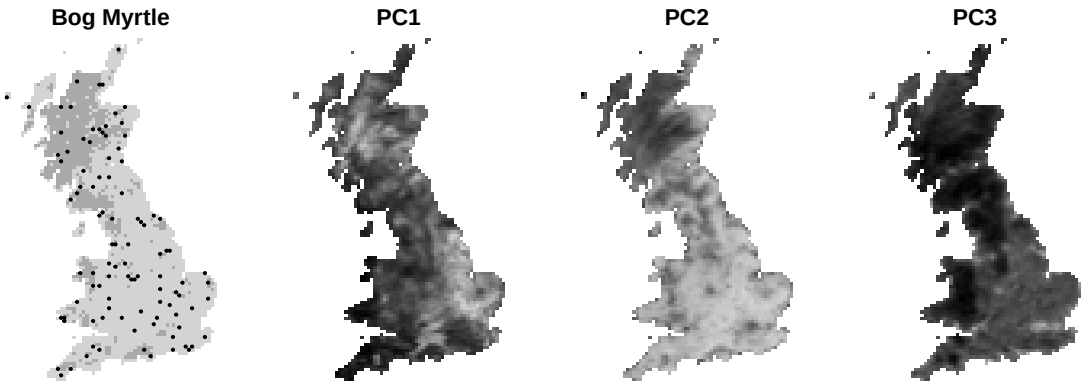
We load these packages and data:

```
> library(GRaF)
> library(raster)
> load(BogMyrtleData.RData)
```

`BogMyrtleData.RData` contains three objects: `PCs` - a `RasterBrick` object containing 10 Principal Components environmental layers; `BM` - a `RasterLayer` object with the known distribution of Bog Myrtle at 10km resolution, and `points` - a `SpatialPoints` object (from the `sp` package, loaded by `raster`) with the 300 randomly selected grid squares used in the model comparison.

We only use the first 100 of these random datapoints and the first 3 environmental layers. We plot the random points over the known distribution along with the environmental covariates:

```
> par(mfrow = c(1, 4), mar = c(1, 1, 2, 1), oma = c(2, 0, 2, 0))
> greys <- colorRampPalette(c(light grey, black))
> plot(BM, col = c(light grey, dark grey), axes = F, box = F,
         legend = F, main = Bog Myrtle)
> plot(points[1:100], add = T, pch = 21, cex = 0.3)
> plot(PCs, 1, col = greys(50), axes = F, box = F, legend = F)
> plot(PCs, 2, col = greys(50), axes = F, box = F, legend = F)
> plot(PCs, 3, col = greys(50), axes = F, box = F, legend = F)
```



To fit the models we extract the environmental data and occurrences for the 100 random points:

```
> y <- extract(BM, points[1:100])
> x <- extract(PCs[[1:3]], points[1:100])
> x <- data.frame(x)
```

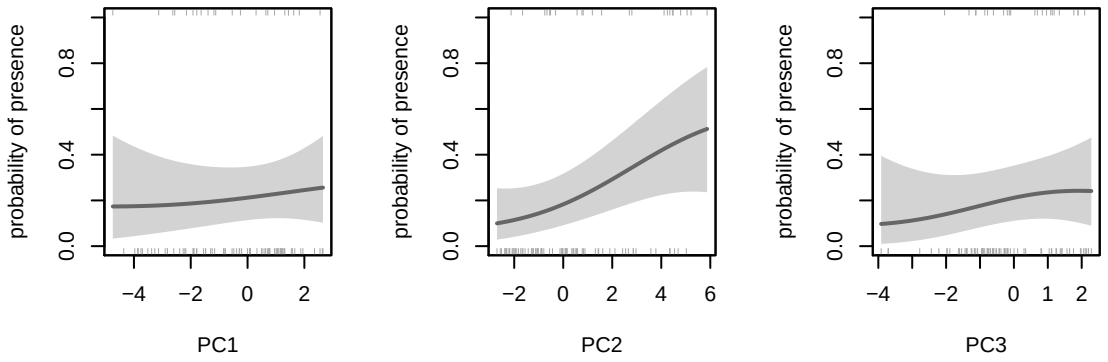
Fitting and visualising GRaF models

To begin with we fit a `graf` model to this data without tuning the hyperparameters by setting `opt.l = FALSE`. When we do this `graf` makes a guess at appropriate hyperparameters from the data. Details of how it does this are given in the helpfile, which can be accessed using `?graf`.

```
> m1 <- graf(y, x, opt.l = FALSE)
```

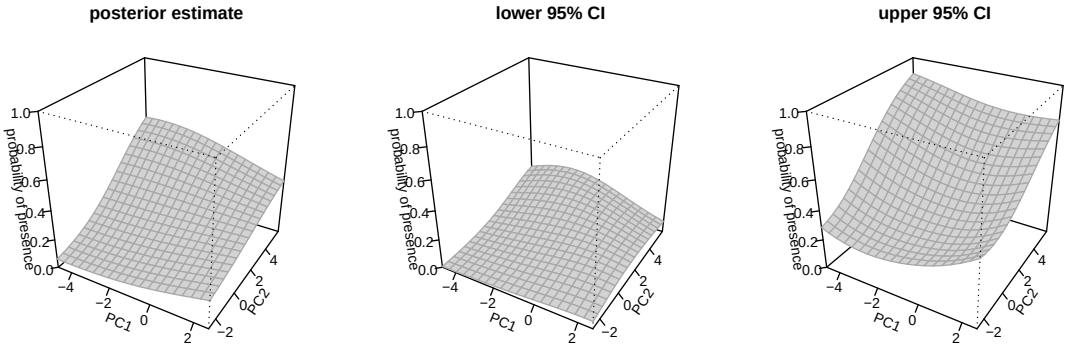
We set up the plotting panel and use `graf`'s `plot` function to visualise the fitted random field along each of the covariates, along with 95% credible intervals and the presence-absence data shown on the axes. See `?plot.graf` for more plotting options.

```
> par(mfrow = c(1, 3))
> plot(m1)
```



We visualise the interaction between the first two covariates with a 3-dimensional perspective plot by using the `plot3d` function:

```
> par(mfrow = c(1, 3))
> plot3d(m1, c(1, 2), prior = FALSE)
```



Note that we suppress plotting of the mean function which in this case defaults to a flat surface.

Model complexity

We can calculate the deviance information criterion (DIC) of the model using the `DIC` function.

```
> DIC(m1)
       DIC         pD
99.387577  4.605478
```

`DIC` also returns the number of *effective parameters* of the model (pD). This is a measure of the complexity of the model and is directly analagous to the number of

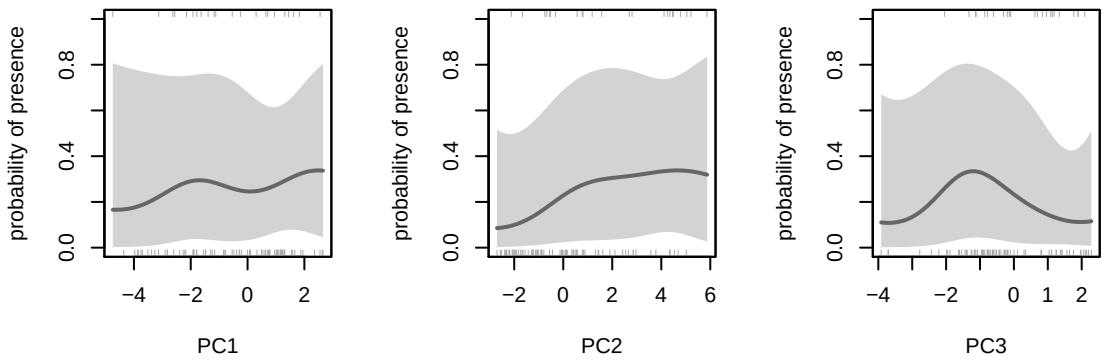
parameters in standard statistical models such as logistic regression.

We can adjust the complexity of the model by setting the lengthscale hyperparameters using the `l` argument. We access the lengthscales from our first model and create a model with much shorter lengthscales (a more complex model) and a model with much longer lengthscales (a less complex model):

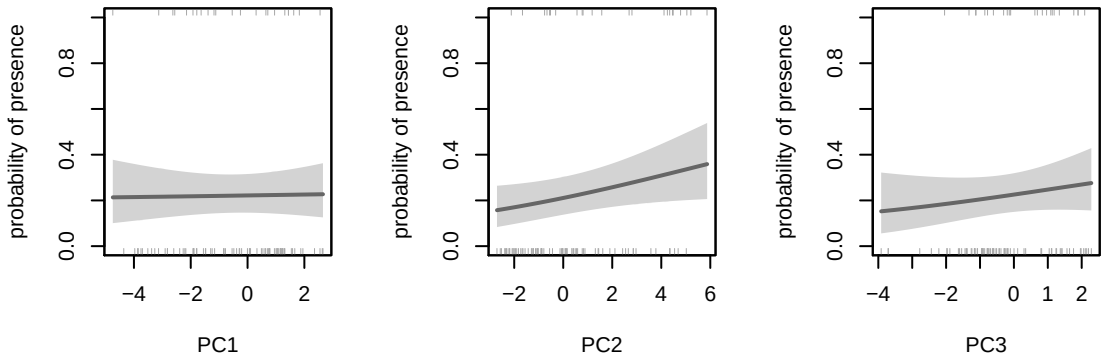
```
> m1$l
[1] 8.069940 8.759522 6.820797
> m2 <- graf(y, x, opt.l = FALSE, l = m1$l * 0.1)
> m3 <- graf(y, x, opt.l = FALSE, l = m1$l * 10)
> DIC(m2)
     DIC       pD
99.96790 13.66754
> DIC(m3)
       DIC         pD
102.476963   2.107745
```

We visualise the surfaces fitted by these models:

```
> par(mfrow = c(1, 3))
> plot(m2)
```



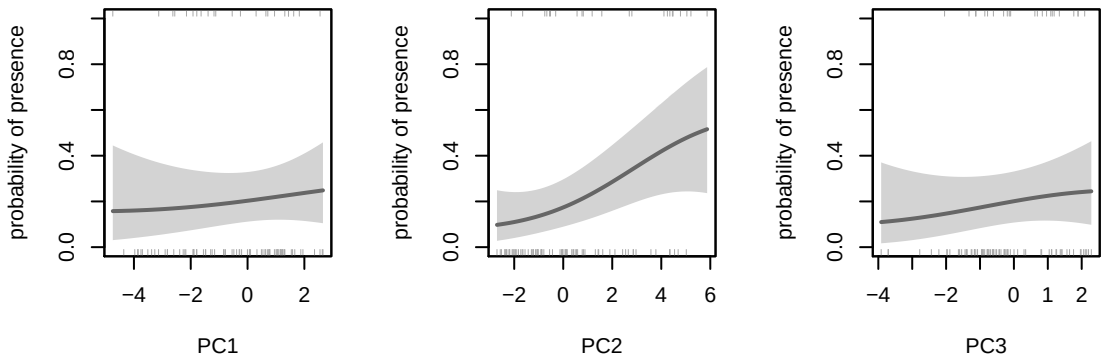

```
> par(mfrow = c(1, 3))
> plot(m3)
```



Optimising the lengthscales

In practice we would prefer to calculate an optimal value for the lengthscales and we can do this by setting `opt.l = TRUE`. `graf` now runs multiple models in an optimisation routine and therefore takes longer to run. Because this optimisation trades complexity off against fit to the data, it guards against model overfitting in a similar way to regularisation methods such as the lasso.

```
> par(mfrow = c(1, 3))
> m4 <- graf(y, x, opt.l = TRUE)
> m4$l
[1] 10.821436  6.769179 11.956630
> DIC(m4)
      DIC        pD
99.029131  4.204211
> plot(m4)
```



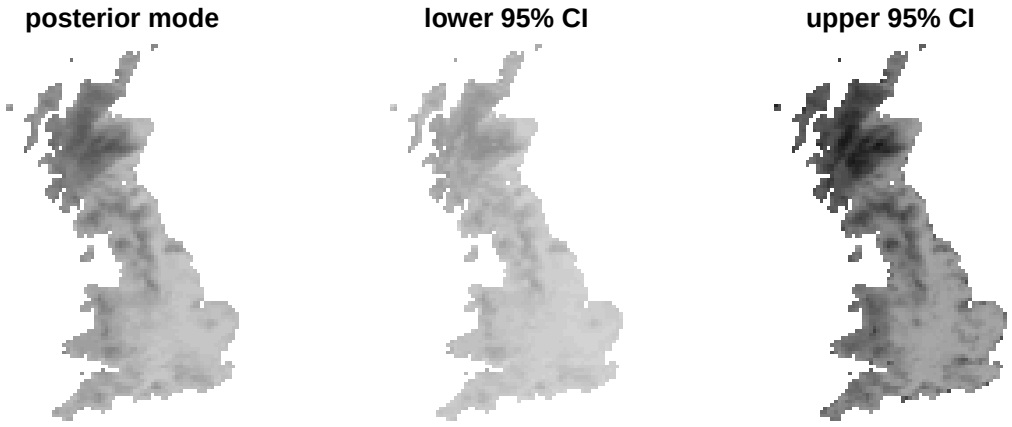
Making predictions

To generate predictions from this model for the rest of Great Britain we extract the relevant environmental covariates and use `graf`'s `predict` function (see `?predict.graf` for details).

```
> unmask <- which(!is.na(PCs[[1]][]))
> xall <- extract(PCs[[1:3]], 1:ncell(PCs))[unmask, ]
> p <- predict(m1, data.frame(xall))
> head(p)
     posterior mode lower 95% CI upper 95% CI
[1,]      0.3590668    0.1934960    0.5569247
[2,]      0.3560689    0.1936734    0.5503185
[3,]      0.3206399    0.1838494    0.4876237
[4,]      0.3027491    0.1671409    0.4730940
[5,]      0.2971340    0.1443506    0.4982656
[6,]      0.2776231    0.1517646    0.4400118
```

As well as a posterior mode - the 'best guess' prediction - lower and upper 95% credible intervals around these predictions are also provided. We can plot these predictions by putting them back into a raster file:

```
> pred <- PCs[[1:3]]
> pred[][unmask,] <- p
> plot(pred, col = greys(50), zlim = c(0, 1), nc = 3,
         box = F, axes = F, legend = F, main = colnames(p))
```

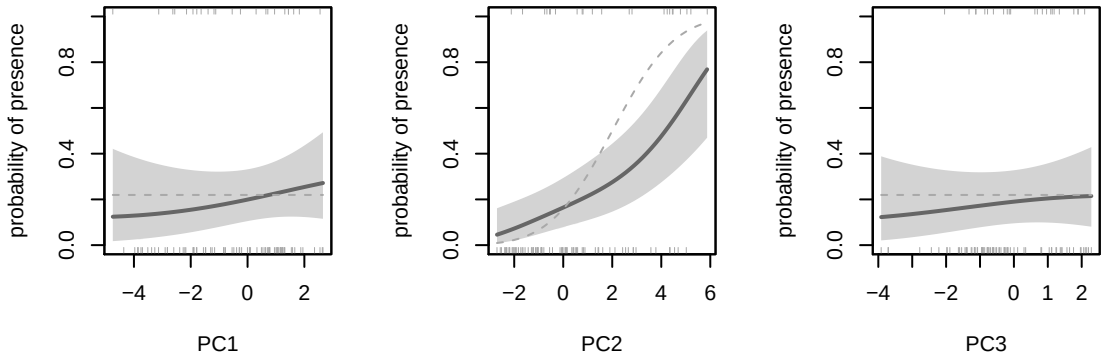


Adding prior ecological knowledge

We can augment the distribution data with knowledge of the ecology of Bog Myrtle. As its name would suggest, Bog Myrtle displays a preference for upland bogs, a habitat with which PC2 is positively correlated (see Appendix B). We specify a simple linear

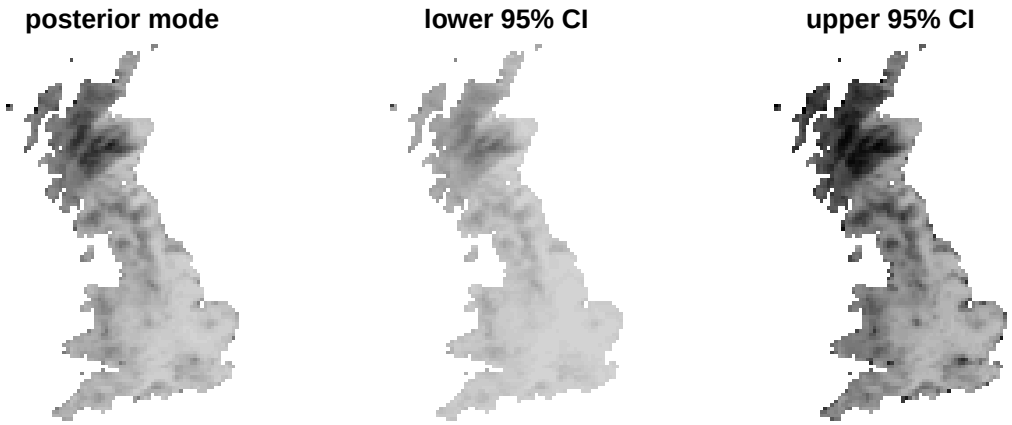
probit model which assigns high probability of presence to sites with a high value of PC2, and low probability of presence elsewhere. We plot the fitted surfaces, with the prior overlaid as a dashed line:

```
> pri <- function(x) pnorm(-1 + 0.5 * x$PC2)
> m5 <- graf(y, x, opt.l = TRUE, prior = pri)
> par(mfrow = c(1, 3))
> plot(m5, prior = T)
```



the predicted surface is a compromise between our prior knowledge of the species' ecology and what can be determined from the data. We plot the predictions from this model:

```
> pred[][unmask,] <- predict(m5, data.frame(xall))
> plot(pred, col = greys(50), zlim = c(0, 1), nc = 3,
        box = F, axes = F, legend = F, main = colnames(p))
```



Chapter 6

High resolution distribution maps of potential vector mosquitoes in the United Kingdom

Nick Golding, Miles A. Nunn & Bethan V. Purse

Authors' contributions: NG collated the occurrence data, devised and carried out the modelling and analysis and wrote the manuscript. MAN and BVP contributed to the design of the study, interpretation of the results and helped to revise the manuscript.

Abstract

A number of vector-borne diseases have recently undergone significant changes in global distribution. Emerging mosquito-borne diseases like West Nile virus (WNV) could be spread in the UK by native mosquito fauna. At present there is limited understanding of which areas of the UK would be at risk from WNV and other diseases if they were to be introduced. Understanding spatial variation in the risk of mosquito-borne diseases requires knowledge of the distributions of the potential vector mosquitoes. Empirical data on these distributions are scarce, so models are required to increase our identify areas of disease risk.

We collate a comprehensive database of occurrence records for UK mosquitoes. Using a novel Bayesian framework we incorporate expert-opinion knowledge to build distribution models for twelve species which account for various sources of bias in the data. The predicted probability of presence of these mosquitoes is mapped at a national-scale as well as the uncertainty in these predictions. Using knowledge of the biology of these species, we map the probability of presence of communities of mosquitoes likely to be capable of sustaining WNV transmission and infecting humans.

These maps are the first high-resolution distribution maps of UK vector mosquitoes and provide a tool for assessing risk from mosquito-borne diseases. The distributions of potential WNV vector communities and of several individual species are positively associated with the distribution of coastal and floodplain grazing marsh. We identify regions of the UK where future field studies and surveillance should be focussed to improve

knowledge of the ecology and distribution of potential vector mosquitoes in the UK.

Introduction

Recent years have seen a number of vector-borne diseases extend their global distributions (Randolph & Rogers, 2010). In some cases pathogens have become established in new areas following the introduction of efficient exotic vectors (Benedict *et al.*, 2007). In others, introduced pathogens have been spread by competent native vector species (e.g. West Nile virus in North America (Hayes, 2001) and Bluetongue virus in Northern Europe (Carpenter *et al.*, 2009)).

The transmission (and therefore the distribution) of vector-borne diseases depends on various factors relating to the pathogen, the vectors and the hosts (Reisen, 2010). Of these, the presence of competent vectors is a pre-requisite for transmission. For many vector-borne pathogens only a single vector species is required to maintain a transmission cycle and infect humans. For zoonotic arboviruses such as West Nile virus (WNV), transmission to humans often requires the presence of multiple vectors. Ornithophagic mosquitoes can sustain transmission of WNV in an enzootic cycle in birds but ornithphagic and mammalophagic 'bridge' vectors are needed for onward transmission to humans (Marra *et al.*, 2004). In the case of WNV and related arboviruses the risk posed by the disease is therefore conditional on the distribution of *communities* of vector species. An understanding of the spatial distributions of potential vectors is therefore fundamental to identifying areas of disease risk.

Whilst no medically important mosquito-borne diseases are thought to be circulating in the UK at present, a number of pathogens could be transmitted by native mosquitoes. These include arboviruses such as West Nile, Sindbis and Tahyna viruses (Medlock *et al.*, 2005, 2007b) and the formerly established malaria-causing *Plasmodia* spp. (Kuhn *et al.*, 2003; Hutchinson, 2004) as well as the less common *Dirofilaria* (Medlock *et al.*, 2007a). Mosquito species vary in their vector competence for these pathogens. The risk posed to the UK by these pathogens, both now and in the future, is therefore dependent upon the distributions of multiple mosquito species (Snow & Medlock, 2006)

Previous studies have investigated this risk by considering the distributions of occurrence records for potential vectors (Snow *et al.*, 1998; Medlock *et al.*, 2007b). However the scarcity of records and a clear bias in recording effort limit the interpretability of these data. In order to better identify areas at risk, robust continuous distribution maps of UK mosquitoes are needed.

Species distribution models (SDMs) have been successfully applied to map the distributions of medically important vector mosquitoes from national (Van Bortel *et al.*, 2009; Kulkarni *et al.*, 2010) to global scales (Benedict *et al.*, 2007; Sinka *et al.*, 2010a,b) as well as the diseases they transmit (Rogers *et al.*, 2002; Bhatt *et al.*, 2013). In their most widely used form, SDMs model the probablity of presence of the species at a site as a function of the environmental conditions at that site. By making predictions from these models to a set of gridded environmental variables covering the area of interest, SDMs can be used to generate maps of the species' expected distribution.

Two studies have used SDMs to model the distributions of anopheline mosquitoes in the UK (Kuhn *et al.*, 2002; Sinka *et al.*, 2010a). These studies generated predicted distribution maps for *Anopheles atroparvus* and *An. messeae* (considered potential malaria vectors) at 8.5km and 5km resolutions as part of European-scale distribution models. No maps of the predicted distributions of other mosquito species in the UK are available in the published literature.

The data used to build SDMs often consist of occurrence records collated from a variety of different sources, such as museum collections, ad hoc field surveys and the published literature (Graham *et al.*, 2004). Such data are subject to various sources of error which can lead to inaccurate model predictions. These include sampling bias (since people are more likely to look for mosquitoes in some areas than others, Phillips *et al.* (2009)) and regression dilution (since locations of many occurrence records are imprecise Frost & Thompson (2000)). Perhaps the most fundamental flaw is that most observations provide *presence-only* data, consisting of records from sites where the species has been found rather than presence and absence data arising from planned

surveys. Most SDMs applied to such data may only predict a species-specific index of the relative suitability of a site for the species, rather than an estimate of the absolute probability that the species is present (Ward *et al.*, 2009; Elith *et al.*, 2011; Golding, 2013a). To produce robust maps of the distributions of these species which can be directly compared and combined into communities, we need methods to overcome these issues.

Here, we collate a comprehensive dataset of the occurrence records of mosquito species in the UK. We elicit knowledge of the habitat requirements and prevalence of these species from an expert in their field ecology to augment these data. Using a new Bayesian approach (Golding & Purse, 2013) we combine these two sources of information to build models of the probability of presence for each of these species in the UK. We apply methods to account for error arising from the expert-opinion estimates, imprecise occurrence records and recording bias in the occurrence data. The approach enables us to account for and present uncertainty in our model predictions.

Using these models, we generate high-resolution maps of the probability of presence of twelve mosquito species in the UK and combine them to identify areas where potential WNV vector communities are likely to occur. Interrogating these maps we identify land cover types associated with individual species and with WNV vector communites. Finally, we identify regions of the UK where future field research should be focussed to improve understanding of the ecology and distributions of UK mosquitoes.

We provide GIS layers of the high-resolution distribution maps in the supplementary material to enable further use of our results. We also provide the complete dataset of species occurrences and computer code for the statistical programming language R (R Development Core Team, 2012) to enable others to repeat our analyses.

Methods

Mosquito occurrence data

We collated occurence records for UK mosquito species from a comprehensive search of the published literature and combined these with additional data from online databases, private collectors, research projects and museums. Records dated before the 1st January 1970 were excluded, since these historical records may relate to environmental conditions which are no longer represented by the environmental covariates.

The online publication databases Pubmed and Web of Knowledge were searched to identify publications using terms derived from species names of the known UK mosquitoes as listed in Table 1 of Medlock *et al.* (2005) with additional terms where species names have changed. These search terms are listed in Appendix 6.B, Table 6.B.1. Additional relevant publications were identified from a comprehensive bibliography of British mosquitoes (Snow *et al.*, 1997). Publications were obtained and all post-1970 mosquito records with useable useful geographic information were extracted.

Occurrence records were also collated from the online data portals of the National Biodiversity Network (`data.nbn.org.uk`, which comprises data from the British Mosquito Recording Scheme) and the MosquitoMap project (`www.mosquitomap.org`). Occurrence data were extracted from labelled pinned mosquito collections of the Oxford University Museum of Natural History and the National Museum of Wales. Additional occurrence data from planned surveys and amateur collectors were provided by UK mosquito researchers (Stefanie Schäfer and Jolyon Medlock *pers. comm.*).

Occurrence records for *Ochlerotatus cantans* and *Oc. annulipes* were combined since these two species are very difficult to distinguish morphologically and have similar habitat preferences (Becker *et al.*, 2010). For brevity, we refer to *Oc. cantans/annulipes* as a single species herein.

Culex. pipiens sensu lato comprises two subspecies in the UK; *Cx. pipiens sensu*

stricto and *Cx. pipiens molestus*. These two subspecies are morphologically indistinguishable and have different biology and habitat preferences, though the latter is thought to be extremely rare. We therefore discarded records identified as *Cx. molestus* but refer to the remaining records as *Cx. pipiens* s.l. whilst considering this species' vector role to be that of the vastly more common *Cx. pipiens* s.s.

Similarly *An. maculipennis* s.l. comprises three species in the UK: *An. atroparvus*, *An. messeae* and *An. daciae* (Linton *et al.*, 2005). These three species are very difficult to distinguish morphologically and are best separated on the basis of their habitat preferences - with *An. atroparvus* restricted to coastal marshes and *An. messeae* and *An. daciae* inhabiting a wider variety of habitats with a slight preference for fresher water (Marshall, 1938; Medlock *et al.*, 2007b). We therefore only considered records of *An. atroparvus* and manually removed any records of this species more than two kilometres from saline to brackish waterbodies since these are likely to be mis-identifications.

All occurrence records were maintained in a relational database and manually inspected to remove duplicates. Geographic locations which were given as place names were georeferenced and shapefiles were created for all occurrence records in R using the packages `sp` and `ggmap` (Pebesma & Bivand, 2005; Kahle & Wickham, 2012). The database of occurrence records and associated GIS data are provided as supplementary material.

Environmental data

We used a gridded dataset of ten principal component indices of environmental variation in the UK to construct distribution models. These continuous, orthogonal variables incorporate information on the seasonality of the environment and have very high spatial resolution (grid cells 190m by 190m). The dataset was derived from a time series of satellite imagery by Fourier decomposition and principal components analysis and is detailed in Golding & Purse (2013).

The mosquitoes *An. atroparvus* and *Ochlerotatus detritus* are both associated

with brackish to saline water (Becker *et al.*, 2010) - an environmental factor which is poorly characterised by satellite imagery. Distribution models for these two species included a proxy environmental layer for salinity calculated as square root of distance to the sea in metres. The transformation was applied to reflect the rapid decline in salinity moving away from the coast and uniform freshwater conditions further inland.

Gaussian random field models

We constructed latent Gaussian random field (GRF) species distribution models using the R package `GRaF` (Golding, 2013b). GRFs have recently been proposed for SDM (Vanhatalo *et al.*, 2012) and the GRaF approach has been shown to have high predictive power Golding & Purse (2013). GRaF automatically penalises overly complex models and prevents overfitting to the data. As a result, computationally expensive stepwise selection and cross-validation routines required for other approaches were not needed (Steyerberg *et al.*, 1999; Elith *et al.*, 2008). Because GRaF fits Bayesian models efficiently using accurate numerical approximations, we were able to generate large-scale predictions at very high resolution for several species and automatically estimate uncertainty in these predictions.

Presence-background modelling and prevalence estimates

The available national-scale distribution data consist only of sites where the species have been found so standard presence-absence distribution models cannot be directly applied. We therefore applied GRaF in a presence-only framework by augmenting this data with background records and estimates of the prevalence of each species. By correcting presence-background distribution models using prevalence estimates, we generated predictions of the *absolute* probability of presence of each species, rather than the maps of relative probability of presence which are produced by standard presence-background approaches (Elith *et al.*, 2006; Ward *et al.*, 2009).

We carried out this correction using the calibration-corrected naïve approach demonstrated by Golding (2013a) using regression weights. Other available presence-

only methods have likelihood surfaces which are not log-concave of the likelihood surface, precluding the use of a number of model fitting procedures (Phillips & Elith, 2011). The calibration-corrected approach does maintain log-concavity and can be used in conjunction with GRaF. GRaF models fitted to presence-only data using this method have previously been shown to outperform the commonly used MaxEnt presence-only model (Phillips *et al.*, 2006) on a large test set of 10km grid scale plant distributions (Golding & Purse, 2013).

Unfortunately no data are available from which to calculate prevalence of mosquito species at a national scale. We therefore elicited estimates of species' prevalences from expert-opinion using the discrete-habitat prevalence elicitation procedure proposed by Golding (2013a) with the augmented land cover map for England described below. In this study prevalence is defined as the number of the UK grid squares occupied by a species, where the UK contains approximately 6.7 million grid squares, each measuring 190m by 190m (surface area 0.036km^2).

As well as the expert's 'best guess' at each species' prevalence, we obtained an estimate of the expert's uncertainty (upper and lower bounds) in these prevalences (see Appendix 6.E). We incorporated this uncertainty into model predictions by fitting several models with different prevalence estimates and applying an efficient numerical integration procedure to combine predictions. The resulting maps incorporate the full distribution of prevalence estimates and incorporate the different possible prevalences into the prediction uncertainty. The model fitting and numerical integration procedure is detailed in Appendix 6.C.

Recording bias

Occurrence data were subject to substantial spatial bias in recording effort, with clusters of occurrence points corresponding to the locations of surveys or amateur collectors as well as a general bias toward south-east England (Fig. 6.2.1a). When used in combination with randomly-selected background samples, spatial bias in occurrence records can lead to severely biased model predictions (Phillips *et al.*, 2009).

We corrected for this bias by sampling background records under the same spatial bias as occurrence records (Phillips *et al.*, 2009). We interpolated occurrence records for all species using a Gaussian kernel smoother to produce a continuous surface of recording bias (Fig. 6.2.1b, R code provided in supplementary material) A standard deviation of 30km was used in the kernel, representing a compromise between the spatial scale of survey data (i.e. the areas covered by local surveys) and the general bias towards the south and east of the UK. We then randomly sampled grid squares, weighted according to this model of recording bias, to generate a population of 1,000 background occurrence records (Fig. 6.2.1c).

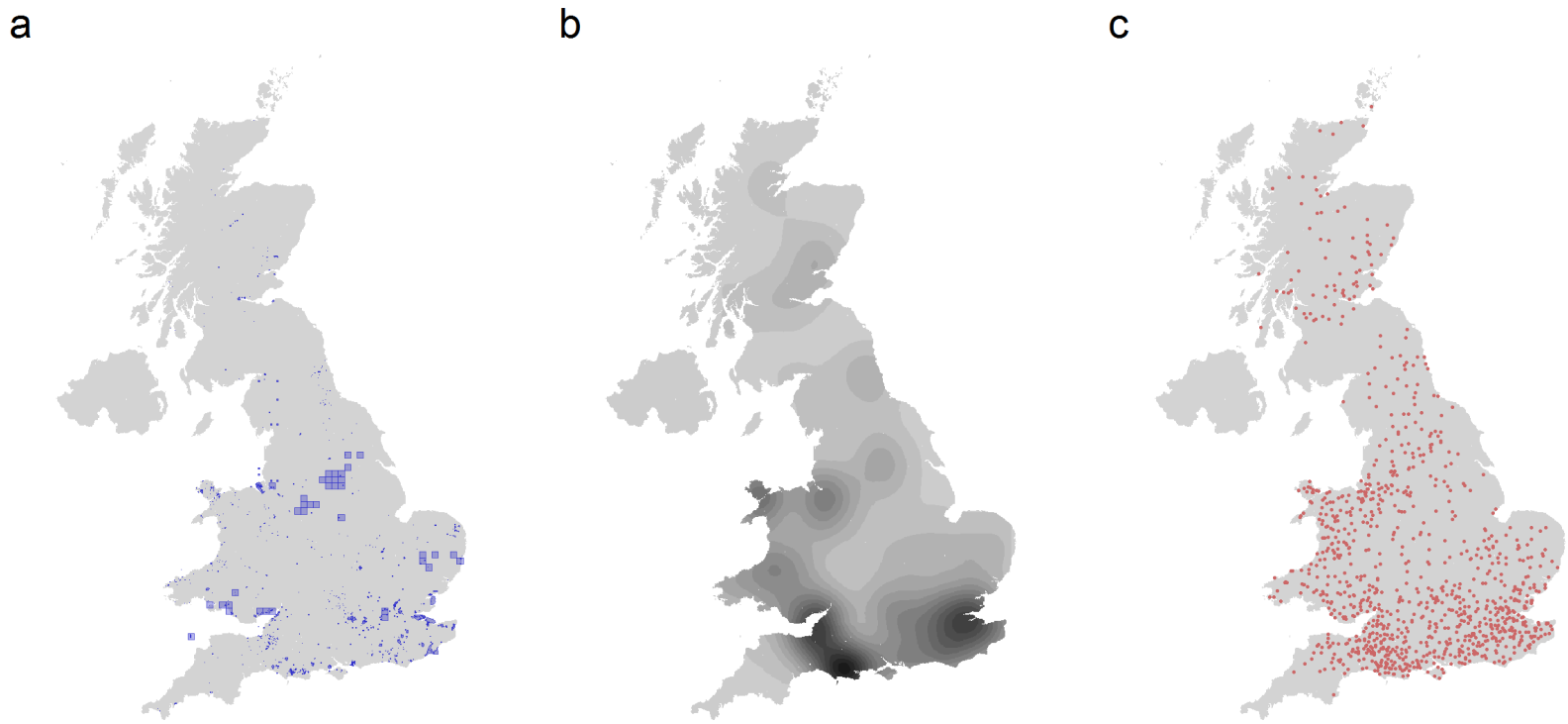


Fig. 6.2.1: **a)** Spatial distribution of occurrence records of varying spatial precision for all mosquito species. **b)** Model of recording bias: an interpolated map of record density for all mosquito species calculated using a Gaussian moving-window smoother. Darker areas represent a higher recording bias. **c)** Location of the 1,000 background samples used to fit distribution models.

Adult dispersal

Whilst our aim was to understand the distribution of the adult mosquitoes which transmit pathogens, the distributions of mosquito larvae are likely to be more strongly constrained by environmental conditions. We therefore produced maps of the probability of presence of *larvae* of each species. Since the flight range of most mosquitoes is greater than the 190m grid cells we consider here, it is not clear whether recorded adult mosquitoes emerged at larval sites in that grid cell or in adjacent cells.

Average dispersal distances are difficult to estimate for adult mosquitoes and consequently very little information is available in the literature. Becker *et al.* (2010) estimate average dispersal distances for adult *Cx. pipiens* of around 500m, though they consider this species to have a relatively short flight range. Wherever specimens were collected as adults or where the life stage was not recorded we increased the area of the occurrence record by adding a buffer zone of 1km

Imprecise record locations

The spatial location of many of the occurrence records was uncertain, either because the location was recorded imprecisely or because it related to an adult or mosquito of unknown life stage. Uncertainty in the true location of occurrence records becomes problematic when extracting corresponding values of environmental covariates with which to fit the SDM. A commonly used approach is to take the average values of environmental covariates in the region of uncertainty. Whilst straightforward to implement, applying an SDM to these average values but not accounting for the associated uncertainty causes *regression dilution* (Frost & Thompson, 2000), which can bias the SDM and its predictions (McInerny *et al.*, 2011). We calculated both means and standard deviations of covariate values for each imprecise record. These statistics were calculated using a weighted procedure, with the weight assigned to each grid cell given by the proportion of that cell falling within the area of the occurrence record. We then fitted GRaF models accounting for this uncertainty

using a measurement-error model (Golding & Purse, 2013). This prevented regression dilution and allowed us to incorporate location uncertainty into model predictions.

WNV community map

We combined single-species distribution maps to identify regions where potential WNV vector communities are likely present. The potential vector status of each of the species is determined on the basis of biting preferences, as given by Medlock *et al.* (2007b).

A potential WNV vector community was defined as one where either a species is present which could fulfill both bridge and enzootic vector roles (*Culiseta morsitans*), or where a potential enzootic vector (*Culex pipiens* s.l.) and at one potential bridge vector (of *Aedes cinereus*, *Coquillettidia richiardii*, *Cs. annulata*, *Ochlerotatus cantans/annulipes*, *Oc. detritus* and *Oc. punctor*) is present. The remaining species that were modelled: *An. atroparvus*, *An. claviger*, *An. geniculatus* and *Oc. rusticus* are considered to be predominantly mammalophilic and are unlikely to play a role in WNV transmission, though they may transmit other diseases and cause biting nuisance.

We estimated the posterior probability of presence of a potential WNV vector community at each grid cell using a Monte Carlo simulation procedure, drawing 1,000 random samples of the probability of presence of each species from the posterior predictive distributions from single species distribution models. The resulting distribution map represents the predicted probability of presence of potential WNV vector communities, as well as uncertainty in this prediction.

Analysing model predictions

Whilst the 'black-box' modelling approach and derived environmental indices we use increase the predictive power of distribution models (Elith *et al.*, 2006; Dormann *et al.*, 2008; Golding & Purse, 2013), the ecological relationships implied by the fitted models are difficult to interpret. We therefore interrogated predicted distribution

maps both by visual comparison with satellite imagery and statistical comparison with land cover data.

The visual inspection aimed to identify characteristics associated with each species' distribution at fine spatial scales. For each species, areas assigned low prediction uncertainty and high or low probability of presence by distribution models were identified. High resolution satellite imagery (Google maps, `maps.google.com`) of these areas was then obtained and inspected to identify fine-scale characteristics of these sites.

Broader-scale correlates of species' distributions were investigated by comparison with the UK Land Cover Map 2007 (LCM2007, Morton *et al.* (2011)) which provides a comprehensive map of land cover types for the UK at high resolution (25m grid cell size). Unfortunately this map does not separately classify coastal and floodplain grazing marsh - an important habitat for a number of UK mosquito species - instead including these habitats in the much broader 'rough grassland' category. We overcome this drawback by augmenting LCM2007 with independent geographic information on the distribution of this habitat type and two others (coastal vegetated shingle and coastal sand dunes) which were also poorly characterised. Whilst these two habitat types are unlikely to be used by mosquitoes, they were included to increase the accuracy of the prevalence elicitation method (described above) which performs better with more land cover classes (Golding, 2013a). We added these habitat types to the gridded LCM2007 data using shapefiles available from Natural England (at `http://www.gis.naturalengland.org.uk`). Where these three habitats were present, they were used to overwrite the LCM2007 classes in the gridded land cover map. Since the shapefiles covered only England and equivalent information for the rest of the UK was unavailable, the resulting land cover map could only be used to assess land cover in England.

We sampled 10,000 190m grid cells at random in England with sampling probability determined by the recording bias map described above. For each grid cell we compared the maximum *a posteriori* probability of presence from distribution maps

with the percentage cover of each land cover class calculated from the 25m resolution land cover map. We screened the 338 potential relationships by calculating spearman's rank correlation coefficients between the probability of presence and percentage cover of each land cover type. For potential relationships with absolute correlation coefficients greater than 0.1 we fitted univariate probit regression models with probability of presence as the response and percentage cover of each land cover type as the covariate. We identified possible relationships by carrying out likelihood ratio tests of statistical significance on these models. To maintain a global error rate at the standard value of 0.05 under multiple comparisons, we adjusted the significance threshold using the correction of Šidák (1967).

Results

Occurrence data and prevalence estimates

A total of 2,993 occurrence records for 21 mosquito species were compiled, representing 1,741 unique locations. No mosquito records were obtained from Northern Ireland or the Scottish Hebrides. Maps of occurrence records for each species are given in Figs. 6.B.2 and 6.B.2 in Appendix 6.B.

Predicted distribution maps were produced for twelve species: *Aedes cinereus*, *Anopheles atroparvus*, *An. claviger*, *Coquillettidia richiardii*, *Cx. pipiens* s.l., *Culiseta annulata*, *Cs. morsitans*, *Oc. cantans/annulipes*, *Oc. detritus*, *Oc. geniculatus*, *Oc. punctor* and *Oc. rusticus*. Species for which distribution maps were not produced either had fewer than 50 occurrence records (*Cx. torrentium*, *An. plumbeus*, *Oc. caspius*, *An. algeriensis*, *Oc. flavescens* & *Cx. europaeus*), represented a complex of morphologically indistinguishable subspecies with different habitat preferences (*An. maculipennis* s.l., Becker *et al.* (2010); Marshall (1938)) or had occurrence records tightly clustered in very few locations (*Cx. modestus*).

Expert opinion prevalence estimates are given in Table 6.E.1 in Appendix 6.E. The modes of these estimates ranged from 0.0002 (*Ae. cinereus*) to 0.066 (*Cs. annulata*). Uncertainty in these estimates was relatively low, with 95% credible intervals ranging in width from 0.0013 (*Ae. cinereus*) to 0.0494 (*Cs. annulata*)

Distribution maps

Distribution maps for individual species are given in Figs 6.A.1 to 6.A.12 in Appendix 6.A. Full resolution raster images of these distribution maps and associated uncertainty are provided in the supplementary material.

Fig. 6.3.1 shows the distribution of potential WNV vector communities. Adult mosquitoes of these eight species are active in the UK from June to August (Medlock *et al.*, 2007b).

Fig. 6.3.2 shows spatial variation in overall prediction uncertainty for single species

distribution models. The Highlands and Islands north-west Scotland have very high uncertainty, as do other upland areas and the Fenland in eastern England. Note that areas of open water such as Lough Neagh in Northern Ireland have high predictive uncertainty due to their disimilarity from sampled sites and background points, though it is unlikely that mosquito larvae would breed here (Becker *et al.*, 2010).

Visual inspection

Areas with highest predicted probabilities of presence and lowest uncertainty for *Oc. cantans/annulipes*, *Oc. geniculatus* and *Oc. rusticus* were all woodland areas, though predictions for the latter two species were particularly uncertain. Predicted distributions of *Ae. cinereus*, *Cs. morsitans* and *Oc. punctor* were all very patchy, with high probability/low uncertainty areas coinciding with wet woodlands. By comparison with these species, *Cq. richiardii*, *An. claviger*, *Cs. annulata* and *Cx. pipiens* s.l. were all relatively widespread with highest probability of presence in wetland areas and lowest in drier areas. *Cs. annulata* and *Cx. pipiens* s.l. were both correctly predicted to be common in urban and suburban areas, albeit subject to uncertainty. Probability of presence of *Oc. detritus* was predicted to be high, but uncertain, in estuarine areas and low elsewhere whereas the predicted distribution of *An. atroparvus* clearly coincided with areas of coastal marsh.

Potential WNV vector communities were predicted to be relatively widespread, with larger high-probability areas focussed on wetland areas, such as the Somerset Levels and Moors and the Fenlands and smaller clusters in wet woodlands. Large areas of Northern Ireland were also predicted have high probability of presence and these all coincided with comparatively wet, low-lying land. All distribution maps predicted low probability of presence in dryer inland areas.

Land cover analysis

Of the 338 vector-land cover relationships in the initial screening procedure 42 had absolute correlation coefficients greater than 0.1 (Tables 6.D.1 and 6.D.2). When anal-

ysed by probit regression, only five of these were significant at a p-value threshold of 0.0012 (Sidak-corrected threshold for a significance level of 0.05 with 42 comparisons; Table 6.D.3). *An. atroparvus*, *An. claviger*, *Cx. pipiens* s.l., *Cs. annulata* and WNV communities were all more likely to be present in cells with a higher percentage cover of coastal and floodplain grazing marsh (Fig. 6.3.3). The strength of these five relationships was similar, with regression coefficients ranging from 0.842 to 1.009 (Table 6.D.3).

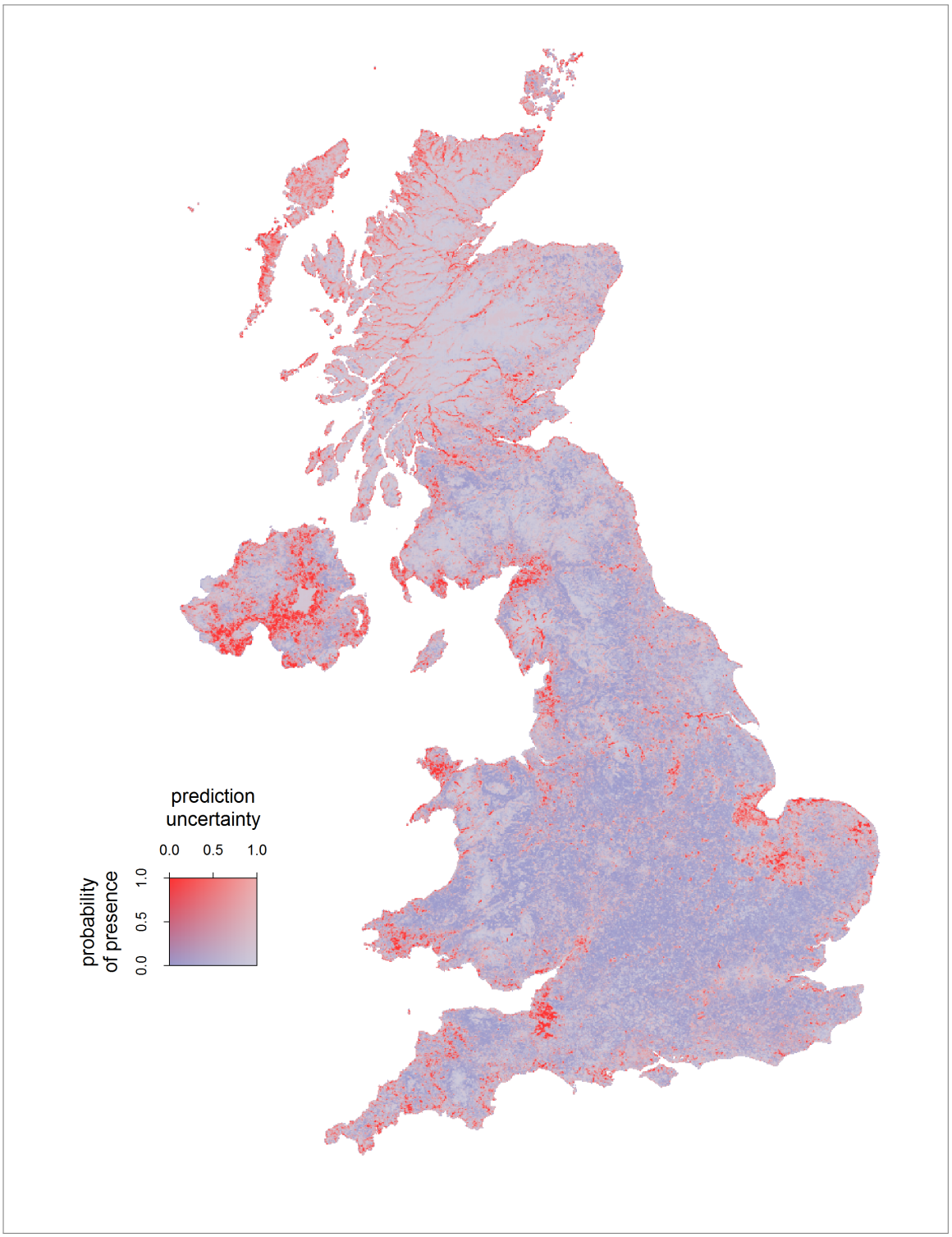


Fig. 6.3.1: Predicted distribution of potential West Nile virus vector communities. Posterior probability of presence of larvae comprising a potential WNV vector community in each 950m by 950m grid cell. The maximum *a posteriori* prediction is shown as a gradient from blue to red and uncertainty in this prediction (width of the 95% credible interval surrounding this estimate) is visualised by the saturation of these colours, with brighter colours denoting low uncertainty and duller colours high uncertainty.

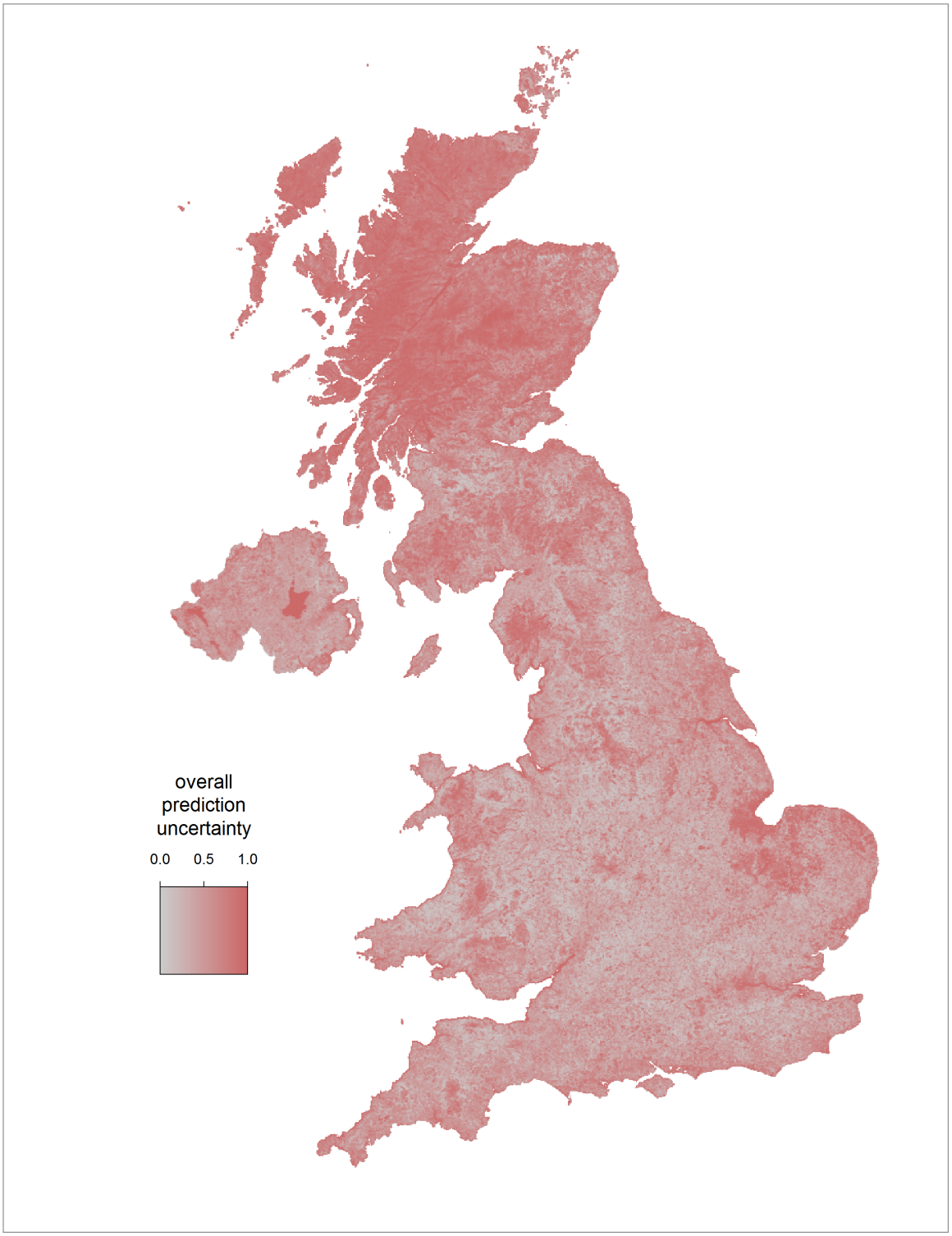


Fig. 6.3.2: Spatial variation in overall prediction uncertainty from the twelve mosquito distribution models. The widths of the 95% credible intervals for each distribution model were summed then scaled to produce an index of uncertainty between 0 (low uncertainty, grey) and 1 (high uncertainty, red) in each 950m by 950m grid cell.

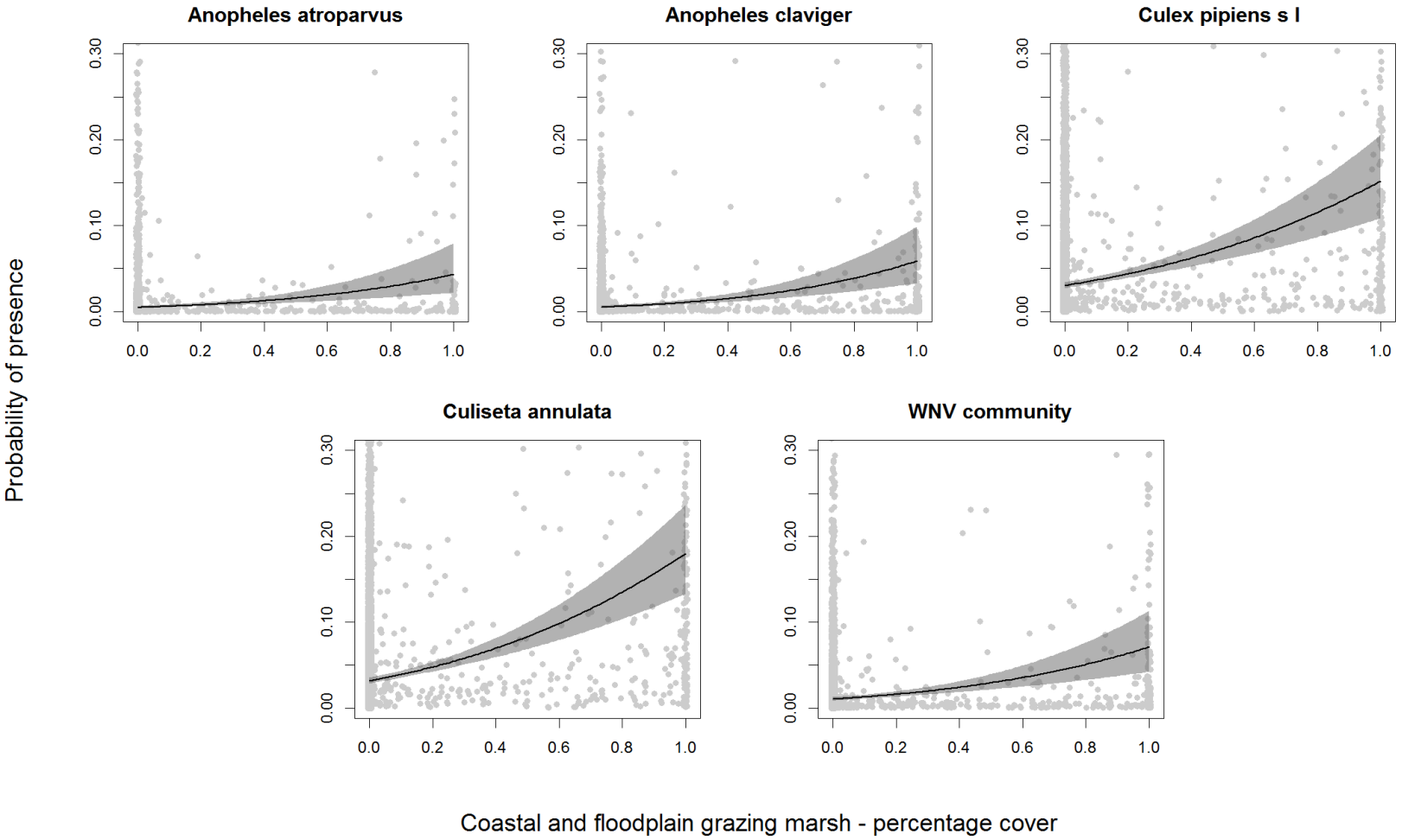


Fig. 6.3.3: Relationship between percentage cover of coastal and floodplain grazing marsh and the probability of presence of four mosquito species and potential WNV vector communities. Probability of presence predicted by probit regression model (black line), 95% confidence interval around this prediction (grey region) and the values of 10,000 evaluation points used to fit these models (grey points, x axis values jittered to aid visualisation)

Discussion

Distribution maps

These are the first published high-resolution distribution maps for potential vector mosquitoes in the UK. Great care was taken to deal with issues associated with the use of the only available data and to honestly express uncertainty in predictions. However due to the paucity of national-scale distribution data for UK mosquitoes it has not been possible to evaluate the accuracy of these maps. We therefore suggest that these maps be treated with caution until higher-quality presence-absence data are available with which to evaluate them.

Improving our understanding of the UK mosquito fauna (particularly in areas such as the Scottish Highlands and Islands which are at present poorly understood) will therefore require comprehensive, broad-scale surveys. Such surveys are likely to be both expensive and time consuming, though similar studies have been carried out successfully at a national scale elsewhere (Van Bortel *et al.*, 2009).

Correlates of species' distributions

Predicted distributions for individual species broadly concurred with their known habitat preferences as described by Marshall (1938) and Becker *et al.* (2010). Our distribution map for *An. atroparvus* differs significantly from those of Kuhn *et al.* (2002) and (Sinka *et al.*, 2010a). The two previous models both predict high relative probability of presence of this species in inland areas of the UK, whereas our map predicts the species to be distributed only in limited coastal areas of the UK. This disagreement may reflect differences in the ecology of *An. atroparvus* in the UK compared with mainland Europe. Whereas the habitat preferences for this species described by Becker *et al.* (2010) for Europe list a variety of different habitats, studies on UK populations have found the species exclusively in coastal marshes (Marshall, 1938; Hutchinson, 2004). It also seems likely that records of *An. atroparvus* used in the previous distribution maps have been confused with the morphologically similar

An. messeae which is found in inland freshwater habitat - an issue which we avoided by careful filtering of occurrence records.

Potential WNV vector communities and a number of individual mosquito species were predicted to be most likely to occur in areas of coastal and floodplain grazing marsh. This land cover type is similar to the coastal marsh habitats in which WNV is found in southern Europe (Ponçon *et al.*, 2007; Danis *et al.*, 2011). The association between wetland areas and mosquitoes is well understood (Becker *et al.*, 2010). Wetlands provide numerous aquatic habitats for larval mosquitoes and are often the epicentre of outbreaks of mosquito-borne disease (Dale & Knight, 2008).

Prevalence-correction

To our knowledge, this is the first study to apply species prevalence estimates to produce robust probability-of presence distribution maps using presence-only data (though Ward *et al.* (2009) tested a related method on a real dataset). By employing these new techniques we were able to combine these distribution maps to map the distributions of potential WNV vector communities. This would not have been possible using standard presence-background SDMs approaches, since since the relative influence of each species on the distribution of potential communities depends on their comparative rarity, information which is not contained in predictions from naïve SDMs.

Mapping ecologial communities

The distribution map of potential WNV vector communities assumed the distributions of the constituent species to be independent of one another. 'Stacking' SDMs in this way is an increasingly common method of modelling the distributions of communities (Ferrier & Guisan, 2006; Benito *et al.*, 2013). Whilst comparatively straightforward to implement, predictions from these models ignore the potential for biotic interactions between species to influence their distributions (Wisz *et al.*, 2013). These approaches may also underpredict co-occurrence between species which respond similarly to en-

vironmental covariates which are missing from the model. Species interaction distribution models have been proposed to deal with these concerns by considering the distributions of multiple species in a single model and explicitly considering correlations in between species' distributions (Kissling *et al.*, 2011). These approaches are still in their infancy, and as yet no methods have been proposed to extend this concept to presence-only records.

Implications for mosquito-borne disease risk

The potential WNV vector community distribution map we provide here provides an indication of which areas of the UK may be most likely to see human WNV cases, should the disease be introduced. This inevitably comes with caveats. These include the inevitable inaccuracies in single-species distribution maps, the potential for biotic interactions to influence the community distribution and specifics of the ecology and biology of species which influence their potential to vector the disease. The map also omits a number of potential vector species, for which insufficient empirical data were available to produce distribution maps. Table 4 of Medlock *et al.* (2007b) lists *Cx. torrentium* and *An. plumbeus* as potential enzootic and bridge vectors respectively and *Cs. litorea* as having the potential to fill both roles.

Cx. modestus is considered the be the most important potential WNV bridge vector present in the UK. This species is considered to be the main bridge vector of WNV in Europe on the basis of its biting behaviour, competence for the virus and implication in outbreaks (Hannoun *et al.*, 1964; Balenghien *et al.*, 2006, 2008). *Cx. modestus* was only recently recognised as being established in the UK and consequently confirmed occurrence records for the species are confined to relatively few locations (Golding *et al.*, 2012; Medlock & Vaux, 2012). In addition, it is unclear whether the species has reached its equilibrium distribution in the UK - a necessity to constructing SDMs on the basis of environmental conditions (Elith & Leathwick, 2009). For this reason it is not yet possible to produce a reliable national-scale map of the distribution of this potential vector. However the species is known to inhabit

wetlands elsewhere in Europe (Becker *et al.*, 2010) and has been found in similar habitat in the UK. The species' equilibrium distribution is therefore likely to coincide to a large extent with the potential WNV vector communities considered here.

In addition to the presence of suitable vectors, potential for the transmission of WNV and other diseases depends on their sufficient abundance and feeding on susceptible hosts at a rate high enough to maintain disease transmission. To more fully understand the risk posed to the UK by WNV and similar diseases, knowledge of the distributions and abundances of potential avian and human hosts is therefore required.

Acknowledgements

We thank Joyon Medlock for providing mosquito occurrence data, Steffi Schäfer for advising on mosquito taxonomy and providing prevalence estimates and David Rogers for helpful comments on the manuscript. We acknowledge funding from the NERC Centre for Ecology & Hydrology (Environmental Change Integrating Fund programme).

References

Balenghien, T., Fouque, F., Sabatier, P. & Bicout, D.J. (2006) Horse-, Bird-, and Human-Seeking Behaviour and Seasonal Abundance of Mosquitoes in a West Nile Virus Focus of Southern France. *Journal of Medical Entomology*, **43**, 936–946.

Balenghien, T., Vazeille, M., Grandadam, M., Schaffner, F., Zeller, H., Reiter, P., Sabatier, P., Fouque, F. & Bicout, D.J. (2008) Vector competence of some French Culex and Aedes mosquitoes for West Nile virus. *Vector-Borne and Zoonotic Diseases*, **8**, 589–95.

Becker, N., Petric, D., Zgomba, M., Boase, C., Madon, M., Dahl, C. & Kaiser, A. (2010) *Mosquitoes and Their Control*. Springer Verlag, Berlin, second edition.

Benedict, M.Q., Levine, R.S., Hawley, W.a. & Lounibos, L.P. (2007) Spread of the tiger: global risk of invasion by the mosquito Aedes albopictus. *Vector-Borne and Zoonotic Diseases*, **7**, 76–85.

Benito, B.M., Cayuela, L. & Albuquerque, F.S. (2013) The impact of modelling choices in the predictive performance of richness maps derived from species-distribution models: guidelines to build better diversity models. *Methods in Ecology and Evolution*, **4**, 327–335.

Bhatt, S., Gething, P.W., Brady, O.J., Messina, J.P., Farlow, A.W., Moyes, C.L., Drake, J.M., Brownstein, J.S., Hoen, A.G., Sankoh, O., Myers, M.F., George, D.B., Jaenisch, T., Wint, G.R.W., Simmons, C.P., Scott, T.W., Farrar, J.J. & Hay, S.I. (2013) The global distribution and burden of dengue. *Nature*.

Carpenter, S.T., Wilson, A.J. & Mellor, P.S. (2009) Culicoides and the emergence of bluetongue virus in northern Europe. *Trends in Microbiology*, **17**, 172–8.

Dale, P.E. & Knight, J.M. (2008) Wetlands and mosquitoes: a review. *Wetlands Ecology and Management*, **16**, 255–276.

Danis, K., Papa, A., Papanikolaou, E., Dougas, G., Terzaki, I., Baka, A., Vrioni, G., Kapsimali, V., Tsakris, A., Kansouzidou, A., Tsiodras, S., Vakalis, N., Bonovas, S. & Kremastinou, J. (2011) Ongoing outbreak of West Nile virus infection in humans, Greece, July to August 2011. *Euro Surveillance*, **16**, 1–5.

Dormann, C.F., Purschke, O., García Márquez, J.R., Lautenbach, S. & Schröder, B. (2008) Components of uncertainty in species distribution analysis: a case study of the Great Grey Shrike. *Ecology*, **89**, 3371–86.

Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M.M., Townsend Peterson, A., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, Robert, E., Soberón, J., Williams, S., Wisz, M.S. & Zimmermann, N.E. (2006) Novel methods improve prediction of species distributions from occurrence data. *Ecography*, **29**, 129–151.

Elith, J. & Leathwick, J.R. (2009) Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 677–697.

Elith, J., Leathwick, J.R., Hastie, T. & R. Leathwick, J. (2008) A working guide to boosted regression trees. *Journal of Animal Ecology*, **77**, 802–13.

Elith, J., Phillips, S. & Hastie, T. (2011) A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, **17**, 43–57.

Ferrier, S. & Guisan, A. (2006) Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology*, **43**, 393–404.

Frost, C. & Thompson, S.G. (2000) Correcting for regression dilution bias: comparison of methods for a single predictor variable. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **163**, 173–189.

Golding, N. (2013a) Methods for eliciting expert-opinion prevalence estimates and incorporating them in presence-only species distribution models. *Manuscript in preparation.*

Golding, N. (2013b) *GRaF: Species distribution modelling using latent Gaussian random fields.* R package version 0.1-0.

Golding, N., Nunn, M.A., Medlock, J.M., Purse, B.V., Vaux, A.G.C. & Schäfer, S.M. (2012) West Nile virus vector Culex modestus established in southern England. *Parasites & Vectors*, **5**, 32.

Golding, N. & Purse, B.V. (2013) GRaF: Fast and flexible bayesian species distribution modelling using gaussian random fields. *Manuscript in preparation.*

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A. (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution*, **19**, 497–503.

Hannoun, C., Panthier, R., Mouchet, J. & Eouzan, J. (1964) Isolement en France du virus West Nile à partir de malades et du vecteur Culex modestus Ficalbi. *Comptes Rendus de l'Académie des Sciences*, **259**, 1.

Hayes, C.G. (2001) West Nile virus: Uganda, 1937, to New York City, 1999. *Annals of the New York Academy of Sciences*, **951**, 25–37.

Hutchinson, R.A. (2004) *Mosquito Borne Diseases in England: past, present and future risks, with special reference to malaria in the Kent Marshes.* Ph.D. thesis, Durham.

Kahle, D. & Wickham, H. (2012) *ggmap: A package for spatial visualization with Google Maps and OpenStreetMap.* R package version 2.1.

Kissling, W.D., Dormann, C.F., Groeneveld, J., Hickler, T., Kühn, I., McInerny, G.J., Montoya, J.M., Römermann, C., Schiffers, K., Schurr, F.M., Singer, A., Sven-

ning, J.C., Zimmermann, N.E. & OHara, R.B. (2011) Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *Journal of Biogeography*, **39**, 2163–2178.

Kuhn, K.G., Campbell-Lendrum, D.H., Armstrong, B. & Davies, C.R. (2003) Malaria in Britain: Past, present, and future. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 9997–10001.

Kuhn, K.G., Campbell-Lendrum, D.H. & Davies, C.R. (2002) A continental risk map for malaria mosquito (Diptera : Culicidae) vectors in Europe. *Journal of Medical Entomology*, **39**, 621–630.

Kulkarni, M.a., Desrochers, R.E. & Kerr, J.T. (2010) High resolution niche models of malaria vectors in northern Tanzania: a new capacity to predict malaria risk? *PLoS ONE*, **5**, e9396.

Linton, Y.m., Lee, A. & Curtis, C. (2005) Discovery of a third member of the Maculipennis group in SW England. *European Mosquito Bulletin*, **19**, 5–9.

Marra, P.P., Griffing, S., Caffrey, C., Kilpatrick, A.M., McLean, R., Brand, C., Saito, E., Dupuis, A.P., Kramer, L. & Novak, R. (2004) West Nile Virus and Wildlife. *BioScience*, **54**, 393.

Marshall, J.F. (1938) *The British Mosquitoes*. Trustees of the British Museum.

McInerny, G.J., Purves, D.W. & McIntyre, K.M. (2011) Fine-scale environmental variation in species distribution modelling : regression dilution , latent variables and neighbourly advice. *Methods in Ecology and Evolution*, **2**, 248–257.

Medlock, J.M., Barrass, I., Taylor, M.A., Kerrod, E. & Leach, S. (2007a) Analysis of climatic predictions for extrinsic incubation of Dirofilaria in the United kingdom. *Vector-Borne and Zoonotic Diseases*, **7**, 4–14.

Medlock, J.M., Leach, S. & Snow, K.R. (2005) Potential transmission of West Nile virus in the British Isles: an ecological review of candidate mosquito bridge vectors. *Medical and Veterinary Entomology*, **19**, 2–21.

Medlock, J.M., Leach, S. & Snow, K.R. (2007b) Possible ecology and epidemiology of medically important mosquito-borne arboviruses in Great Britain. *Epidemiology and Infection*, **135**, 466–482.

Medlock, J.M. & Vaux, A.G.C. (2012) Distribution of West Nile virus vector, Culex modestus, in England. *The Veterinary Record*, **171**, 278.

Morton, D., Rowland, C., Wood, C., Meek, L., Marston, C., Smoth, G., Wadsworth, R. & Simpson, I.C. (2011) Countryside Survey: Final Report for LCM2007 the new UK Land Cover Map. Technical report.

Pebesma, E. & Bivand, R. (2005) Classes and methods for spatial data in R. *R News*, **5**.

Phillips, S.J., Anderson, R.P. & Schapire, Robert, E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J.R. & Ferrier, S. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, **19**, 181–97.

Phillips, S.J. & Elith, J. (2011) Logistic methods for resource selection functions and presence-only species distribution models. *AAAI (Association for the Advancement of Artificial Intelligence*, pp. 1384–1389.

Ponçon, N., Balenghien, T., Toty, C., Baptiste Ferré, J., Thomas, C., Dervieux, A., L'Ambert, G., Schaffner, F., Bardin, O. & Fontenille, D. (2007) Effects of Local Anthropogenic Changes on Potential Malaria Vector Anopheles hyrcanus and

West Nile Virus Vector Culex modestus, Camargue, France. *Emerging Infectious Diseases*, **13**, 1810–5.

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.

Randolph, S.E. & Rogers, D.J. (2010) The arrival, establishment and spread of exotic diseases: patterns and predictions. *Nature Reviews Microbiology*, **8**, 361–71.

Reisen, W.K. (2010) Landscape epidemiology of vector-borne diseases. *Annual Review of Entomology*, **55**, 461–83.

Rogers, D.J., Randolph, S.E., Snow, R.W. & Hay, S.I. (2002) Satellite imagery in the study and forecast of malaria. *Nature*, **415**, 710–715.

Sinka, M.E., Bangs, M.J., Manguin, S., Coetzee, M., Mbogo, C.M., Hemingway, J., Patil, A.P., Temperley, W.H., Gething, P.W., Kabaria, C.W., Okara, R.M., Van Boeckel, T.P., Godfray, H.C.J., Harbach, R.E. & Hay, S.I. (2010a) The dominant Anopheles vectors of human malaria in Africa, Europe and the Middle East: occurrence data, distribution maps and bionomic precis. *Parasites & Vectors*, **3**, 117.

Sinka, M.E., Rubio-Palis, Y., Manguin, S., Patil, A.P., Temperley, W.H., Gething, P.W., Van Boeckel, T.P., Kabaria, C.W., Harbach, R.E. & Hay, S.I. (2010b) The dominant Anopheles vectors of human malaria in the Americas: occurrence data, distribution maps and bionomic precis. *Parasites & Vectors*, **3**, 72.

Smyth, G. (2013) *statmod: Statistical Modeling*. R package version 1.4.17.

Snow, K.R. & Medlock, J.M. (2006) The potential impact of climate change on the distribution and prevalence of mosquitoes in Britain. *European Mosquito Bulletin*, **21**, 1–10.

Snow, K.R., Rees, A. & Brooks, J. (1997) *A revised bibliography of the mosquitoes of the British Isles*. University of East London Press.

Snow, K.R., Rees, A. & Bulbeck, S. (1998) *A provisional atlas of the mosquitoes of Britain*. University of East London Press.

Steyerberg, E.W., Eijkemans, M.J. & Habbema, J.D. (1999) Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *Journal of Clinical Epidemiology*, **52**, 935–42.

Van Bortel, W., Grootaert, P., Hance, T., Hendrickx, G. & Takken, W. (2009) Mosquito Vectors of Disease: Spatial Biodiversity, Drivers of Change and Risk "MODIRISK" Final Report Phase 1. Technical report, Brussels.

Vanhatalo, J., Veneranta, L. & Hudd, R. (2012) Species distribution modeling with Gaussian processes: A case study with the youngest stages of sea spawning whitefish (Coregonus lavaretus L. s.l.) larvae. *Ecological Modelling*, **228**, 49–58.

Šidák, Z. (1967) Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, **62**, 626–633.

Ward, G., Hastie, T., Barry, S., Elith, J. & Leathwick, J.R. (2009) Presence-only data and the em algorithm. *Biometrics*, **65**, 554–63.

Wisz, M.S., Pottier, J., Kissling, W.D., Pellissier, L., Lenoir, J., Damgaard, C.F., Dormann, C.F., Forchhammer, M.C., Grytnes, J.A., Guisan, A., Heikkinen, R.K., Hø ye, T.T., Kühn, I., Luoto, M., Maiorano, L., Nilsson, M.C., Normand, S., Ockinger, E., Schmidt, N.M., Termansen, M., Timmermann, A., Wardle, D.A., Aastrup, P. & Svenning, J.C. (2013) The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biological Reviews of the Cambridge Philosophical Society*, **88**, 15–30.

Appendix 6.A Distribution maps

Figures 6.A.1 to 6.A.12 present the predicted distributions of the twelve potential vector mosquitoes considered in this study. In each figure the probability of presence of at least one larva in each 950m by 950m grid cell is shown as a gradient from blue (low probability) to red (high probability). Prediction uncertainty (measured as the width of the 95% credible interval around the prediction) is shown as a gradient of decreasing saturation, with highly certain predictions having full saturation and uncertain predictions low saturation (grey regions). Full resolution raster images of these distribution maps are provided in the supplementary material.

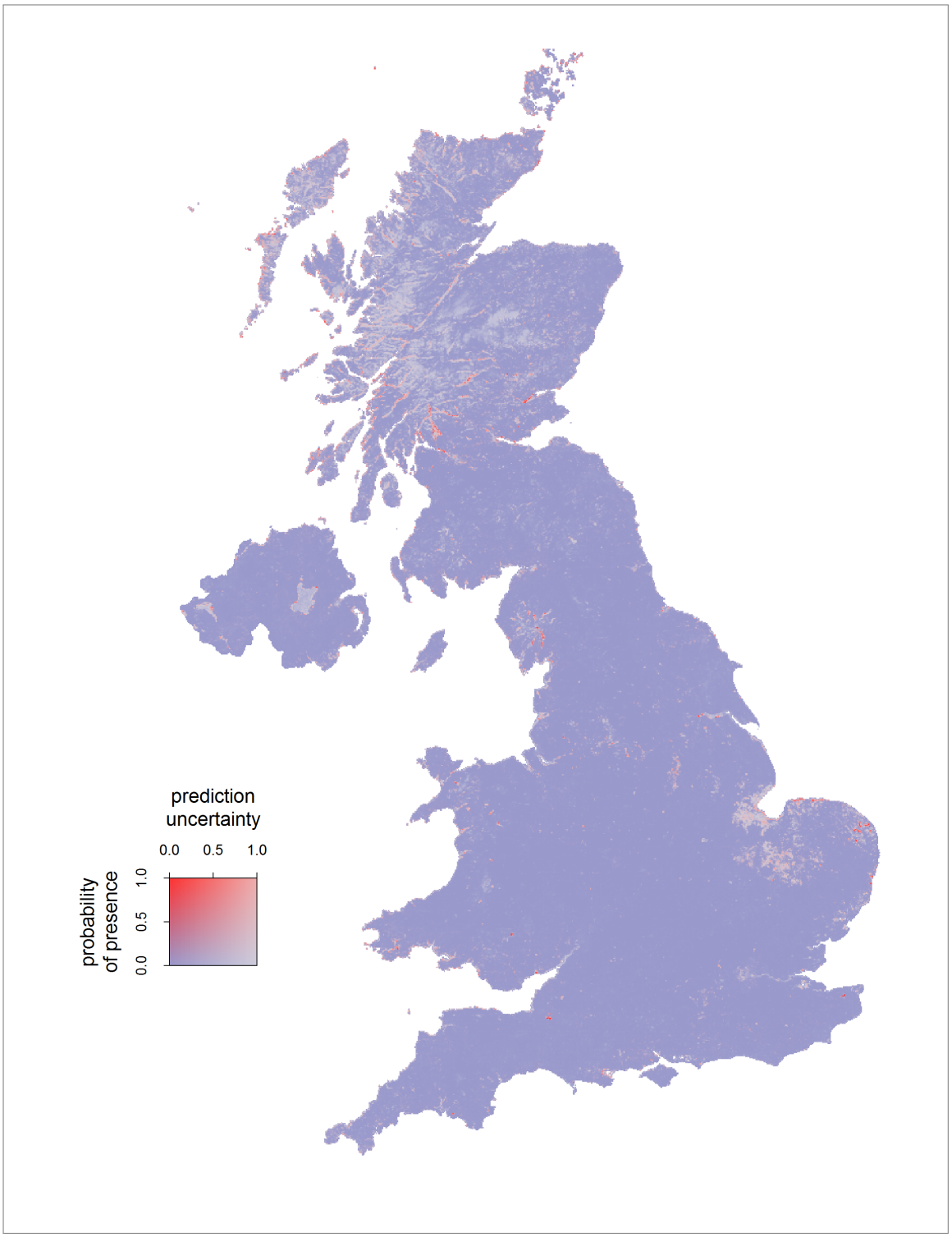


Fig. 6.A.1: Predicted distribution of *Aedes cinereus*.

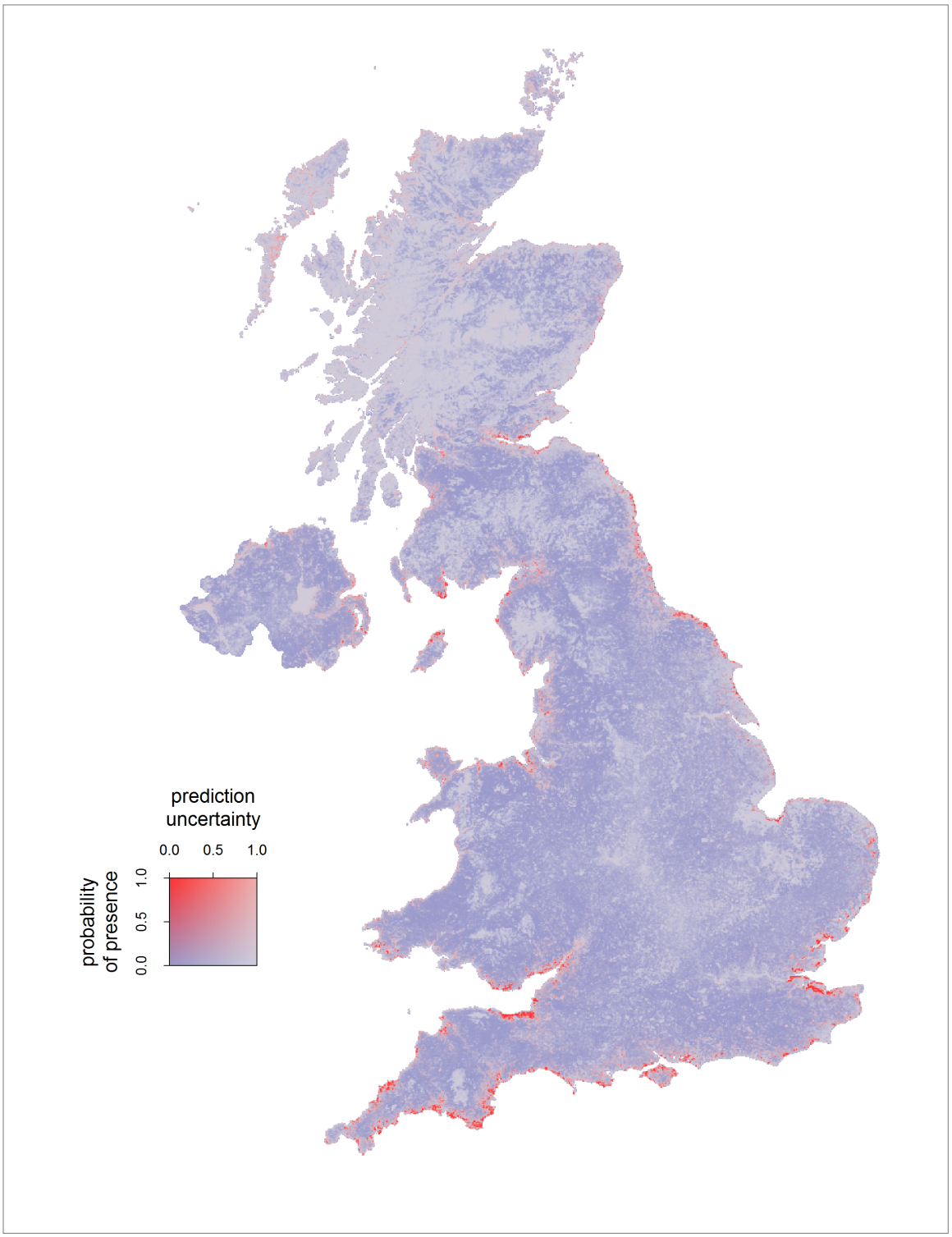


Fig. 6.A.2: Predicted distribution of *Anopheles atroparvus*.

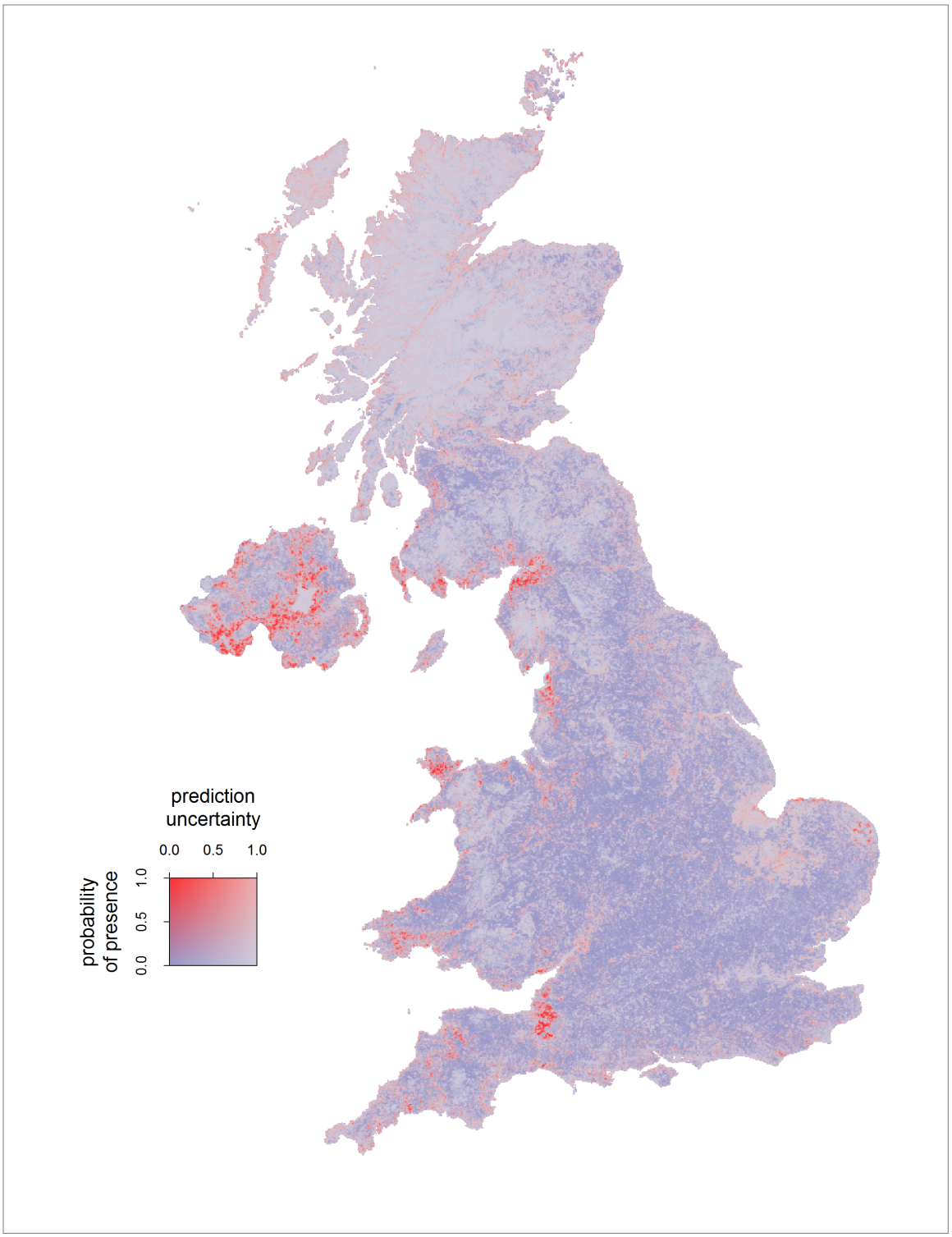


Fig. 6.A.3: Predicted distribution of *Anopheles claviger*.

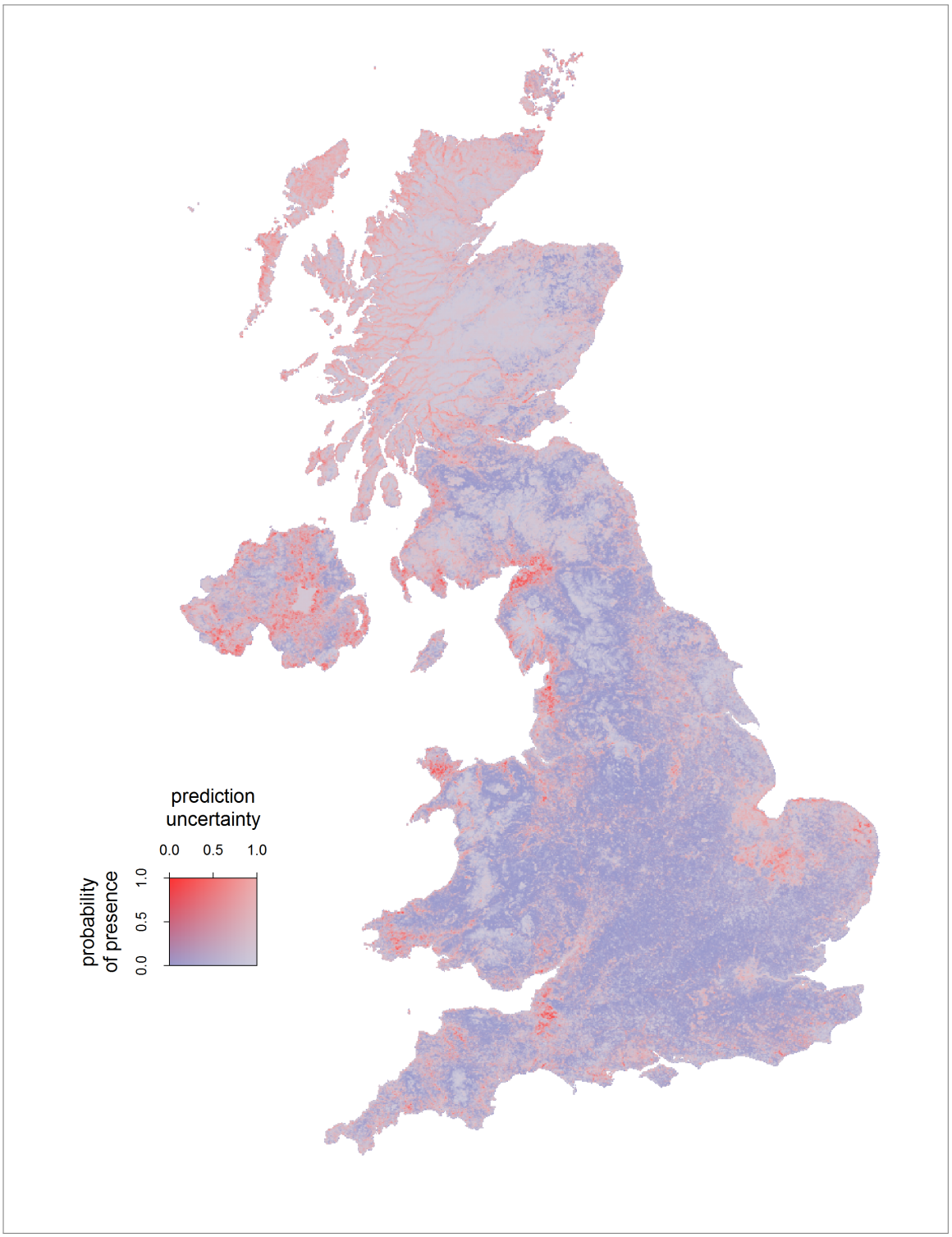


Fig. 6.A.4: Predicted distribution of *Coquillettidia richiardii*.

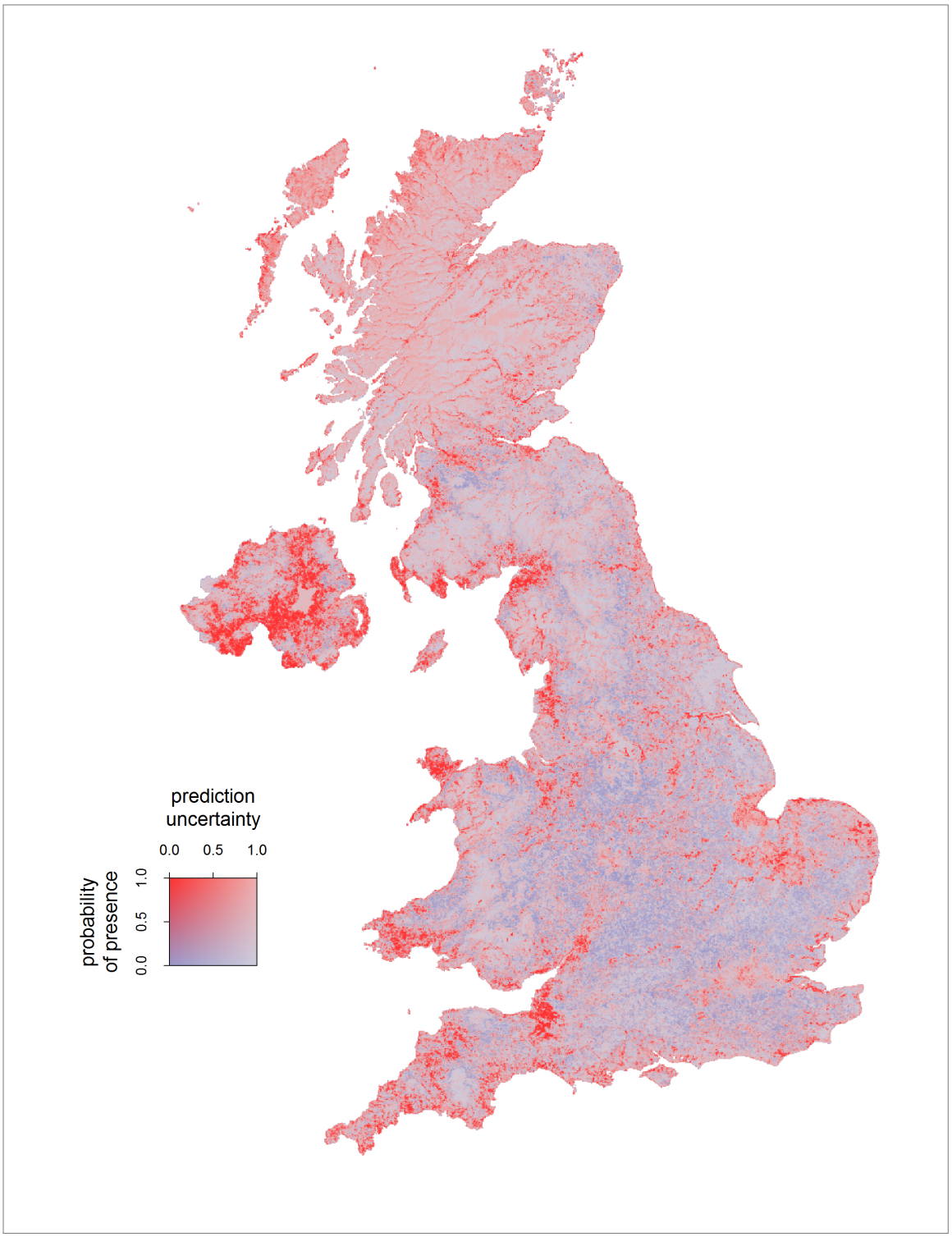


Fig. 6.A.5: Predicted distribution of *Culex pipiens* s.l.

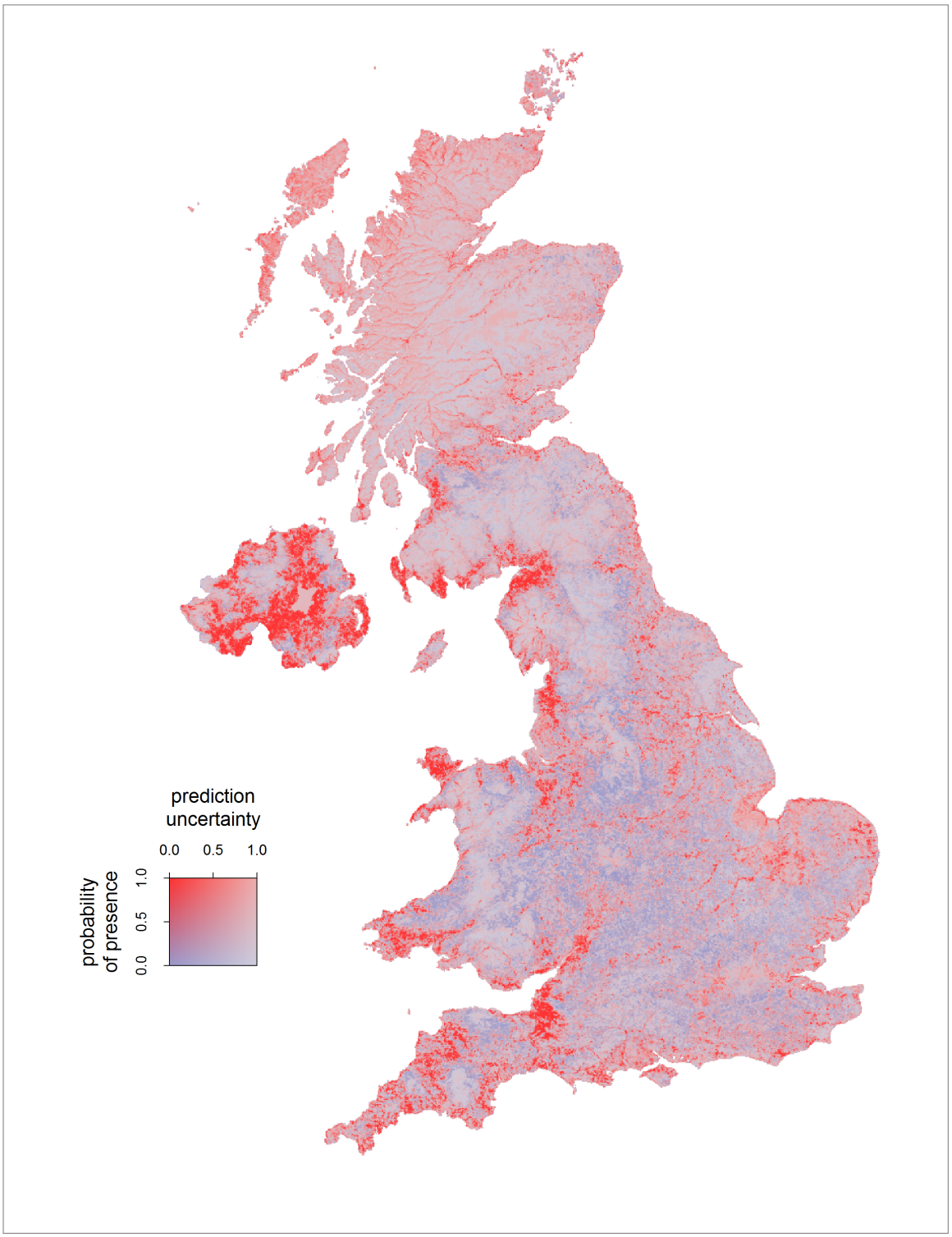


Fig. 6.A.6: Predicted distribution of *Culiseta annulata*.

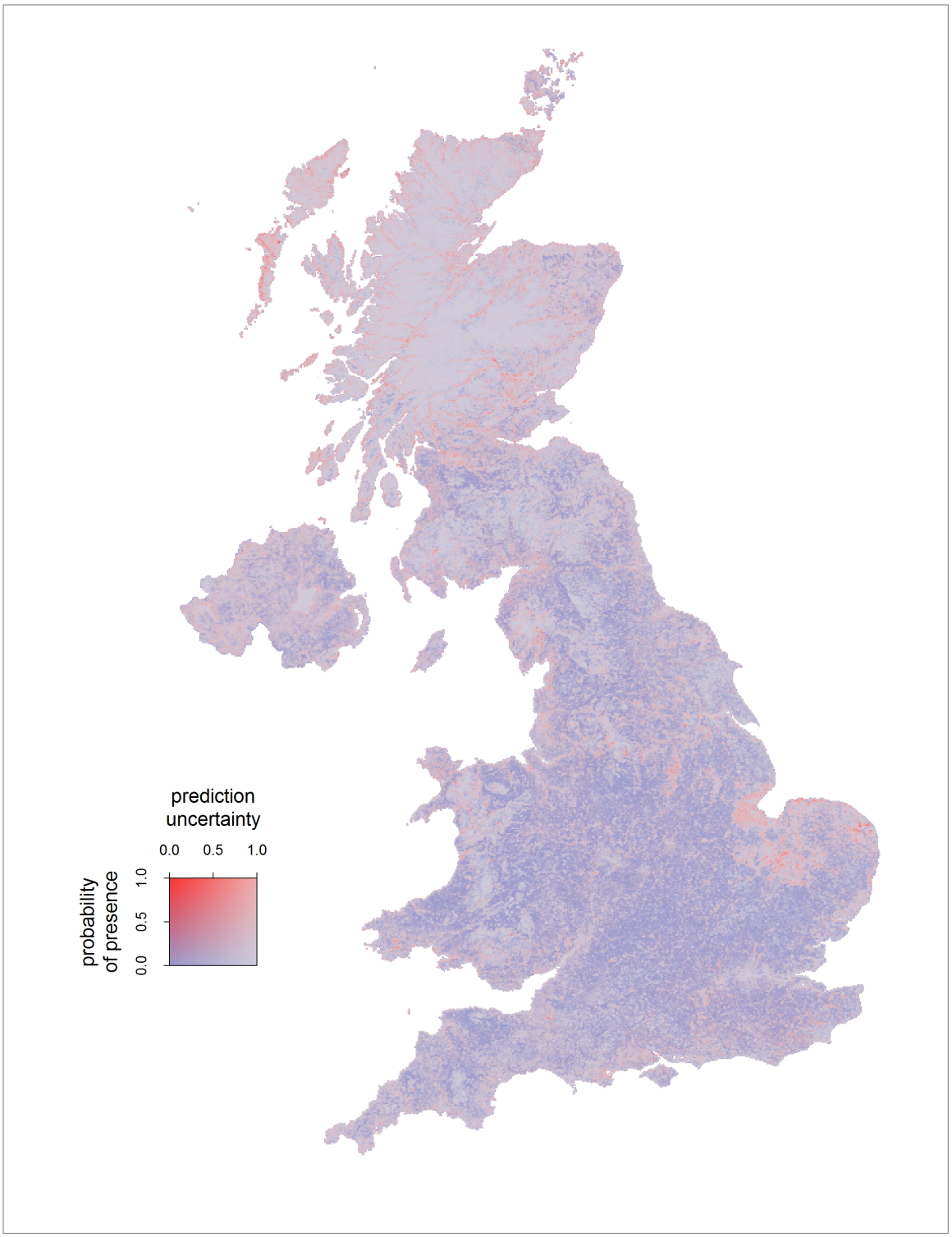


Fig. 6.A.7: Predicted distribution of *Culiseta morsitans*.

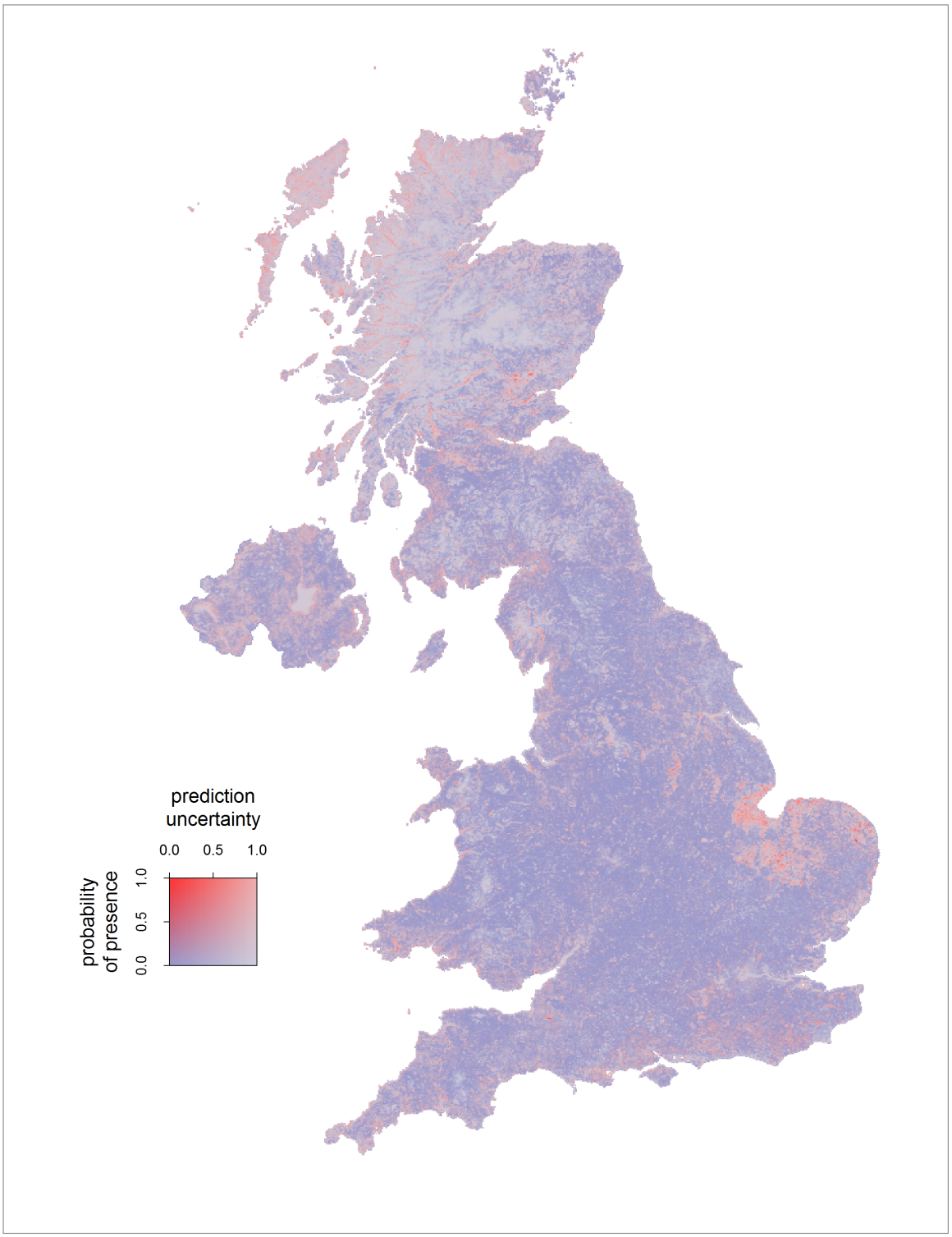


Fig. 6.A.8: Predicted distribution of *Ochlerotatus cantans/annulipes*.

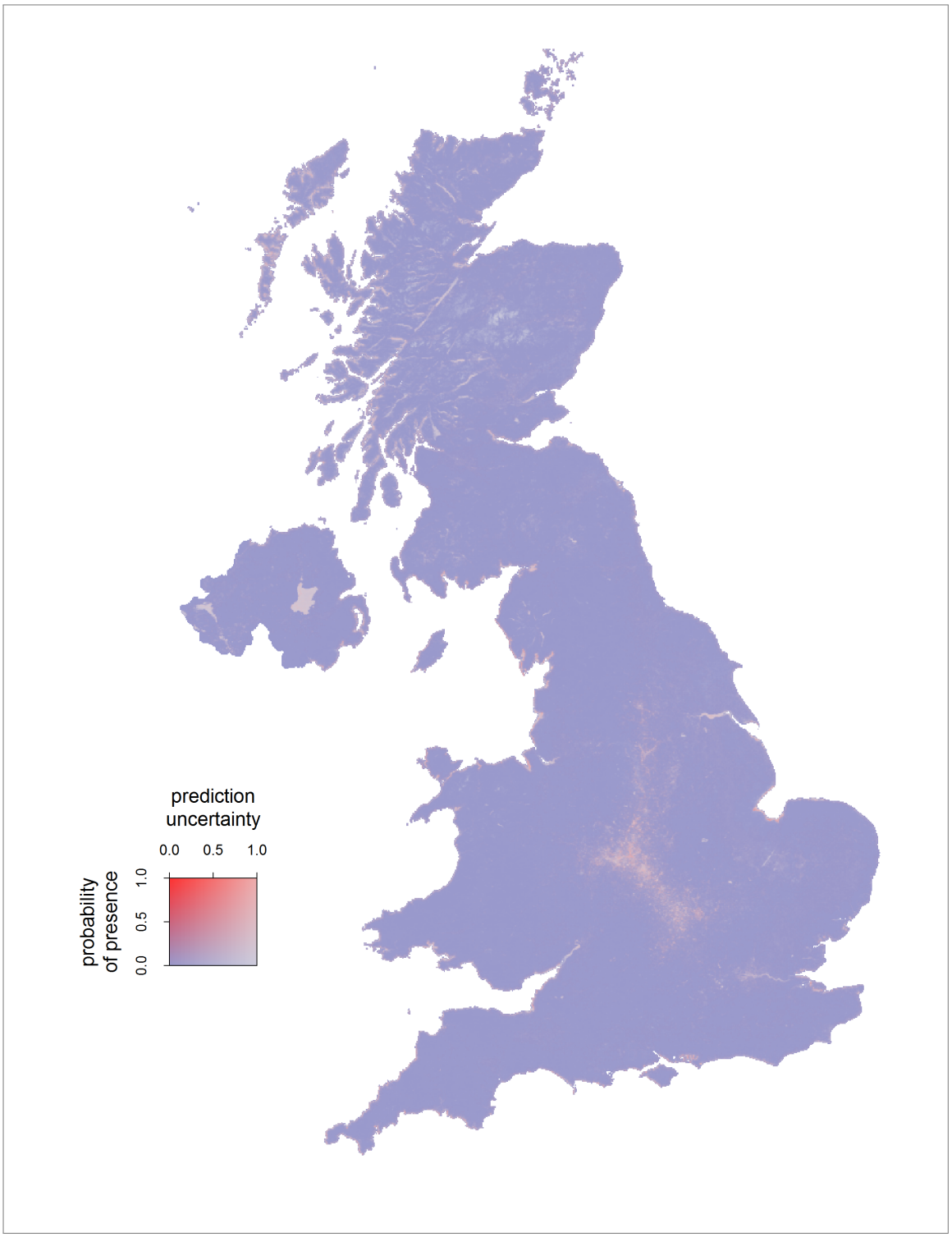


Fig. 6.A.9: Predicted distribution of *Ochlerotatus detritus*.

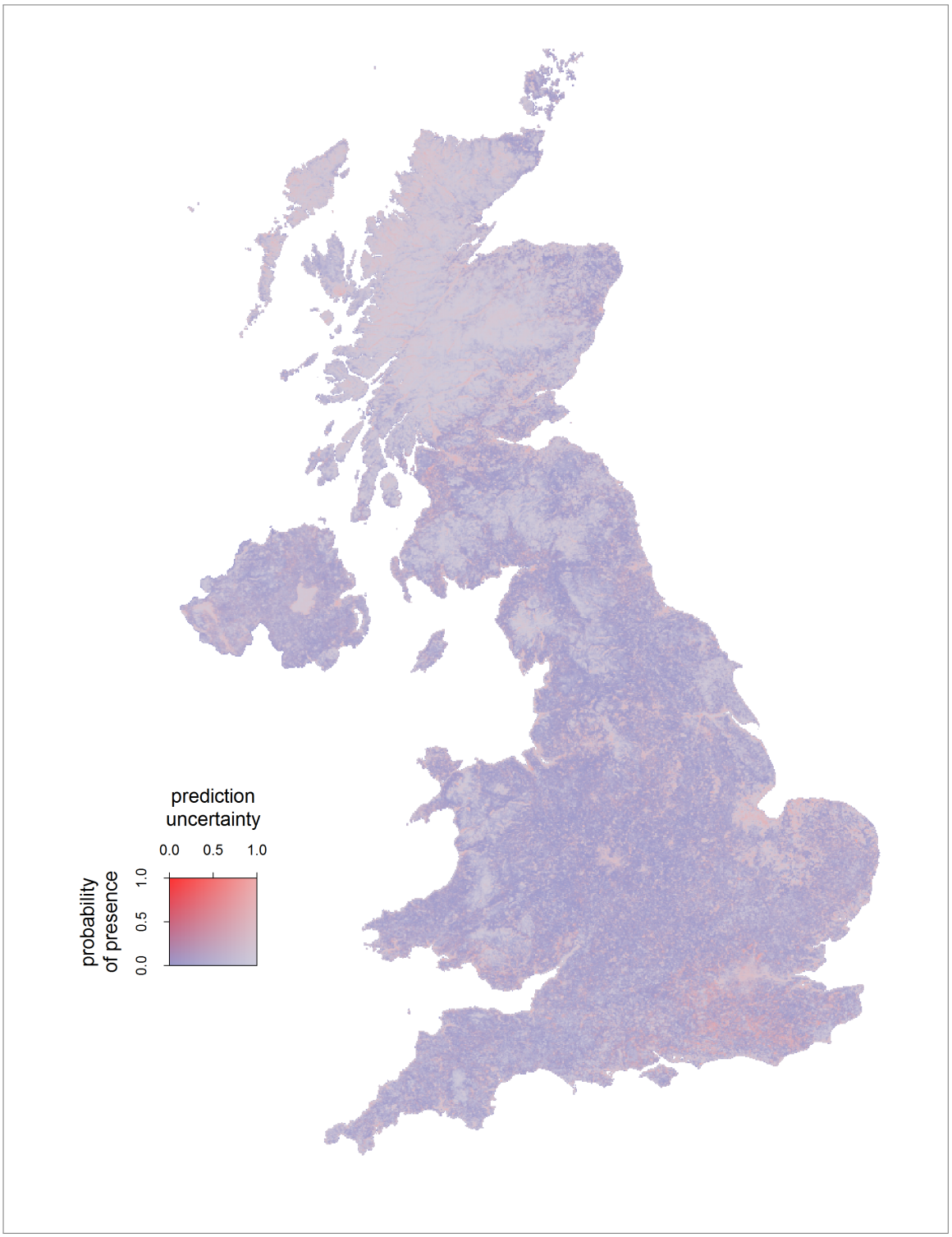


Fig. 6.A.10: Predicted distribution of *Ochlerotatus geniculatus*.

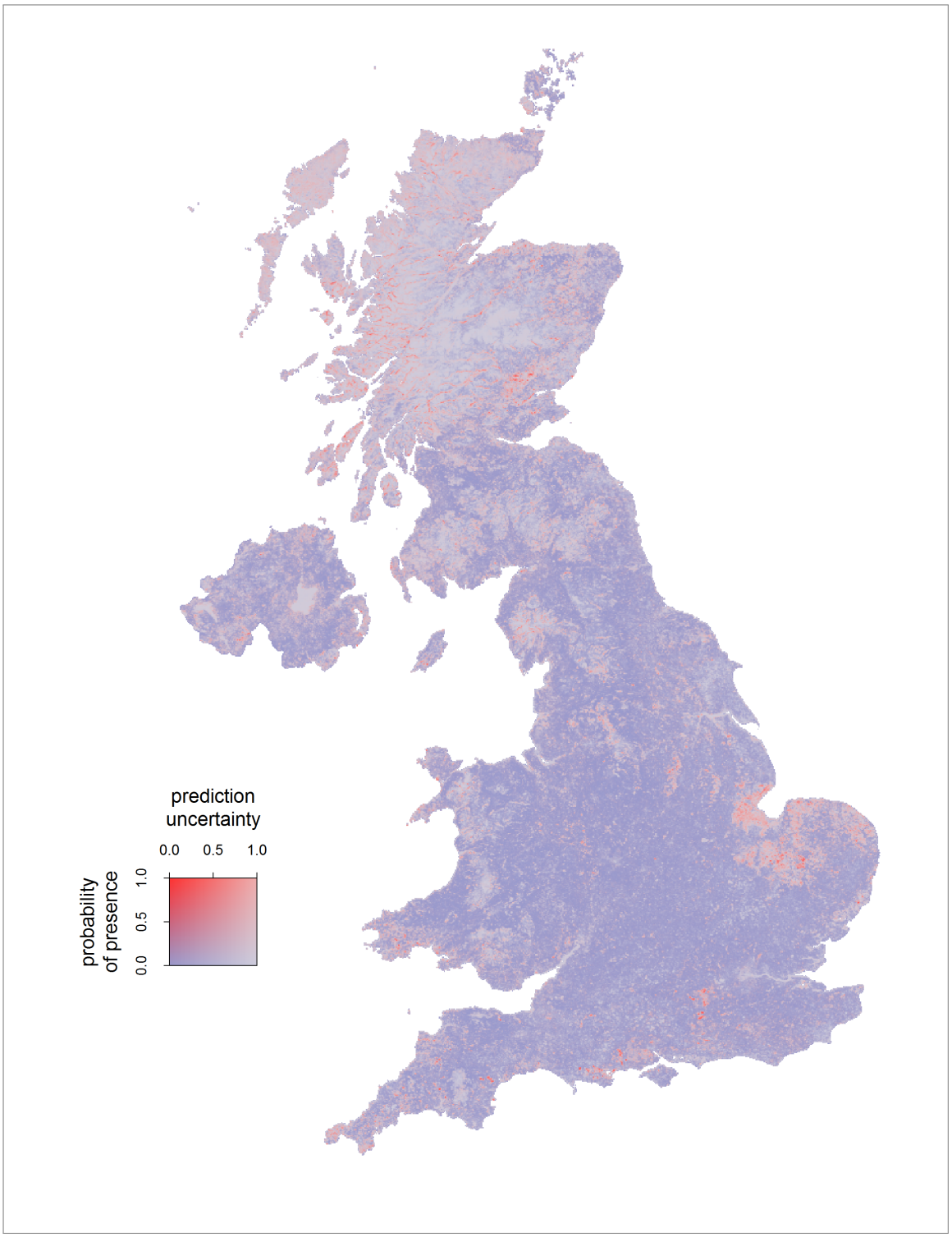


Fig. 6.A.11: Predicted distribution of *Ochlerotatus punctor*.

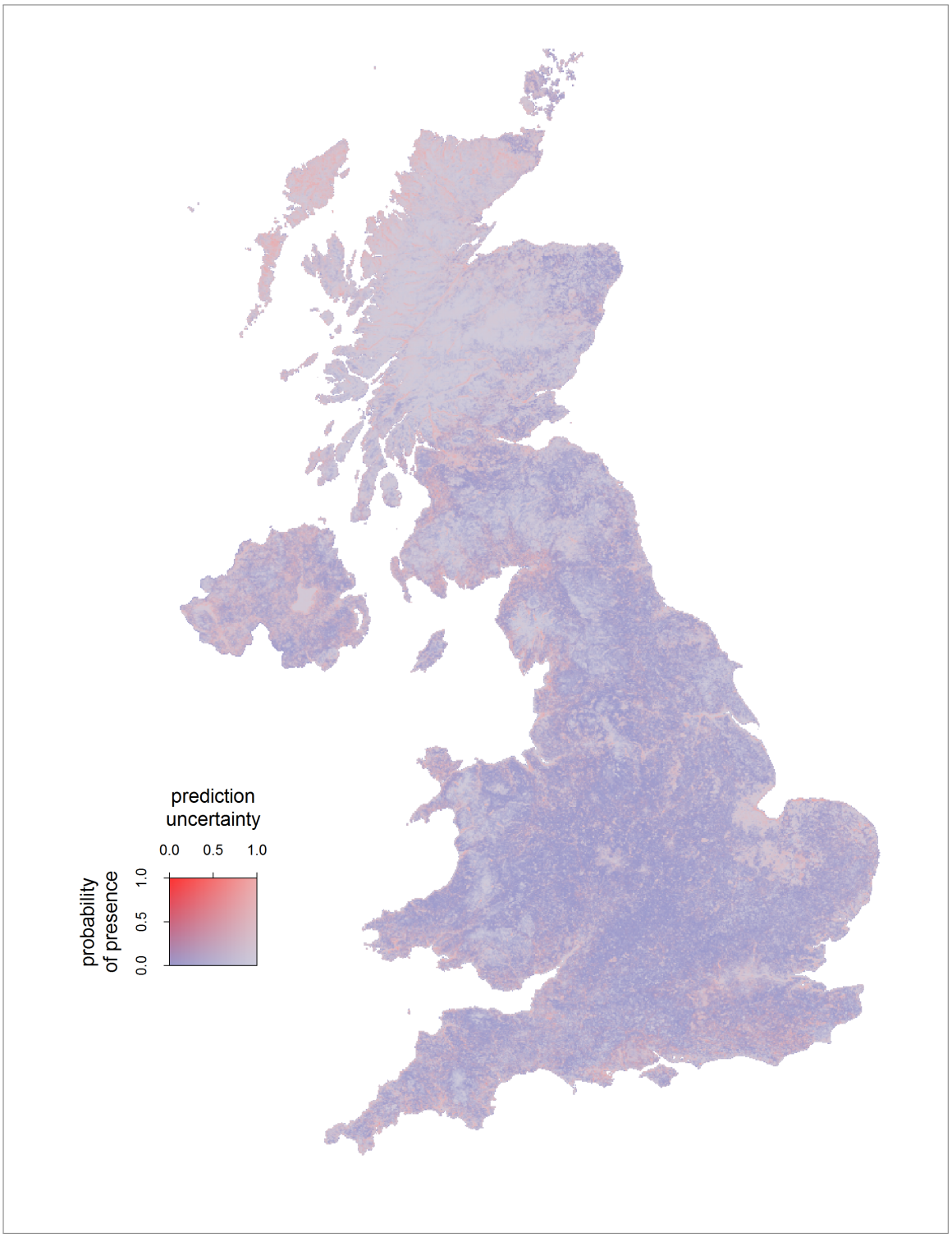


Fig. 6.A.12: Predicted distribution of *Ochlerotatus rusticus*.

Appendix 6.B Mosquito occurrence data

The distribution of dates of occurrence records in the database are given in Fig. 6.B.1. Search terms used in the literature search are given in Table 6.B.1. The spatial distributions of occurrence records for species with at least 20 records are shown in Figs. 6.B.2 and 6.B.3.

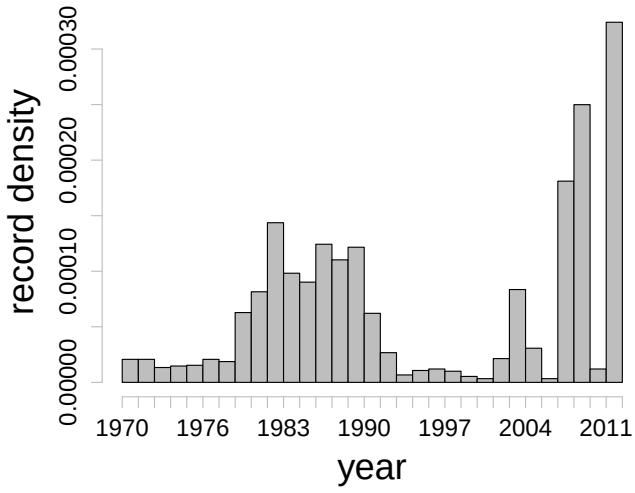


Fig. 6.B.1: Density of occurrence records for all species by date.

Table 6.B.1: Terms used in literature search of mosquito occurrence data. Terms followed by an asterisk are alternative species names.

Aedes cinereus	*Culiseta alaskaensis*	*Ochlerotatus dorsalis*	*Aedes detritus**
Aedes vexans	*Culiseta annulata*	*Ochlerotatus flavescens*	*Aedes dorsalis**
Anopheles algeriensis	*Culiseta fumipennis*	*Ochlerotatus leucomelas*	*Aedes flavescens**
Anopheles claviger	*Culiseta litorea*	*Ochlerotatus punctor*	*Aedes leucomelas**
Anopheles maculipennis	*Culiseta morsitans*	*Ochlerotatus sticticus*	*Aedes punctor**
Anopheles plumbeus	*Culiseta subochrea*	*Ochlerotatus rusticus*	*Aedes sticticus**
Coquillettidia richardii	*Ochlerotatus geniculatus*	*Orthopodomyia pulcripalpis*	*Aedes rusticus**
Culex modestus	*Ochlerotatus annulipes*	*Aedes geniculatus**	*Anopheles atroparvus**
Culex pipiens	*Ochlerotatus cantans*	*Aedes annulipes**	*Anopheles daciae**
Culex torrentium	*Ochlerotatus caspius*	*Aedes cantans**	*Anopheles messeae**
Culex europaeus	*Ochlerotatus communis*	*Aedes caspius**	*Culex territans**
Culiseta longiareolata	*Ochlerotatus detritus*	*Aedes communis**	*Orthopodomyia pulchripalpis**

Fig. 6.B.2: Distribution of occurrence records with lifestage recorded as larvae (blue) or as adults or unrecorded (green). Occurrence records artificially enlarged to aid visualisation.

Fig. 6.B.3: Distribution of occurrence records with lifestage recorded as larvae (blue) or as adults or unrecorded (green). Occurrence records artificially enlarged to aid visualisation.

Appendix 6.C Model fitting and integration of prevalence uncertainty

The species distribution models discussed in this paper use GRaF to fit a latent Gaussian random field (GRF) model to binomial data using a Laplace approximation. Full details of GRaF are given in Golding & Purse (2013). Below we outline our approach of nesting GRaF within a numerical integration procedure to incorporate uncertainty in each species' prevalence estimate into the final prediction.

In the presence/absence case GRaF allows us to fit a latent GRF model and calculate the posterior probability distribution over the probability of presence at a site, conditional on the value of environmental covariates at that site and the Maximum *a posteriori* (MAP) estimate of the GRF hyperparameters θ. In notation:

$$p(y_i = 1|x_i, \hat{\theta}) = \int p(y_i = 1|z_i)p(z_i|\hat{\theta}, x_i)\ \mathrm{d}z_i \tag{6.C.1}$$

where $y_i = 1$ is a binary variable indicating whether site i represents a presence (1) or a background (0) record, x_i gives the value of environmental covariates at site i, $\hat{\theta}$ is the MAP estimate of θ and z_i is the value of the GRF $\mathbf{z}$ at i. GRaF solves this equation efficiently by computing an accurate Laplace approximation to the conditional posterior of $\mathbf{z}$, calculating $\hat{\theta}$ by numerical optimisation and employing an analytical solution to the integral over z_i (which is possible because a probit link is used).

As discussed in the text and in Golding (2013a), this model can be adapted to generate predictions of probability of presence from presence-background data by using a prevalence estimate π to select an appropriate number of background samples. We incorporate uncertainty in this prevalence estimate by specifying a prior probability distribution over π and integrating over this uncertainty giving us the predictive equation:

$$p(y_i = 1|x_i, \hat{\theta}) = \iint p(y_i = 1|z_i)p(z_i|\hat{\theta}, x_i, \pi)p(\pi)\ \mathrm{d}z_i\, \mathrm{d}\pi \tag{6.C.2}$$

Since $p(\pi)$ is one-dimensional, this additional integration can be performed numer-

ically. The prior we specify over the probability estimate follows a beta distribution, which enables us to estimate the integral efficiently using Gaussian quadrature. This procedure is as follows:

1. Select N prevalence values (nodes) π_j and associated quadrature weights w_j to use in the Gaussian quadrature. We select optimal π_j to integrate over the beta prevalence prior distribution with parameters given in Appendix 6.E using the R function `gauss.quad.prob` from the `statmod` package (Smyth, 2013).

2. For each prevalence estimate π_j:

   (a) Calculate the calibration-corrected weighting for the 1,000 background records. nu_j is the expected number of absence records when np presence records are observed assuming prevalence is π_j and is calculated as $nu_j = \frac{np(1-\pi_j)}{\pi_j}$ (Golding, 2013a). Each background point is therefore given a weight of $nu_j/1000$, so that the weights for background records sum to nu_j. The presence records all recieve a weighting of 1.

   (b) Fit a GRaF model with the np presence records and weighted background records and for each grid square i calculate the predicted mean $\tilde{\mu}_{ij}$ and standard deviation $\tilde{\sigma}_{ij}$ of $\tilde{z}_i$ conditional on π_j.

3. For each grid square i:

   (a) Estimate the integrals over $\tilde{\mu}_i$ and $\tilde{\sigma}_i$ as a weighted sum of the nodes and weights:

$$\begin{aligned} p(\tilde{\mu}_i|\hat{\theta}, x_i) &= \int p(\tilde{\mu}_i|\hat{\theta}, x_i, \pi)p(\pi)\,\mathrm{d}\pi \\ &\approx \sum_{j=1}^{N} \tilde{\mu}_{ij} w_j \end{aligned} \tag{6.C.3}$$

$$\begin{aligned} p(\tilde{\sigma}_i|\hat{\theta}, x_i) &= \int p(\tilde{\sigma}_i|\hat{\theta}, x_i, \pi)p(\pi)\,\mathrm{d}\pi \\ &\approx \sum_{j=1}^{N} \tilde{\sigma}_{ij} w_j \end{aligned} \tag{6.C.4}$$

giving an estimate of the posterior distribution of z_i:

$$p(\tilde{z}_i|\hat{\theta}, x_i) = N(\tilde{z}_i; \tilde{\mu}_i, \tilde{\sigma}_i^2) \tag{6.C.5}$$

(b) For each grid square i analytically determine the mode and 95% credible intervals of the posterior distribution over the probability of presence of the species $p(y_i = 1|\hat{\theta}, x_i)$ from $p(\tilde{z}_i|\hat{\theta}, x_i)$ as in (Golding & Purse, 2013).

Gaussian quadrature enables us to calculate accurate approximations to this integral based on relatively few function evaluations. Integration using N nodes will be accurate provided the effect of the prevalence estimate on the probability of presence is well approximated by a polynomial of order $2N$-1. On the basis of a preliminary analysis we set N to 3 (equivalent to a polynomial of order 5) since higher values of N had little effect on predicted probabilities of presence.

For each species we ran this model fitting, prediction and integration procedure to predict the probability of presence of each species for the approximately 6.7 million pixels that make up the UK at the 190m resolution of the environmental covariates. This large computational task was sped up by fitting models and generating predictions in parallel on the NEMESIS computing cluster at the Centre for Ecology & Hydrology, Edinburgh. Using a cluster of around 100 cpus enabled us to carry out the model fitting and prediction for all species in around ten hours.

Whilst these predictions account for uncertainty in the shape of the latent GRF and in the prevalence estimate they do not account for uncertainty in the hyperparameters θ which control complexity of the GRF. Due to the high dimension of θ, standard numerical integration procedures are not feasible and a simulation approach would have beeen computationally prohibitive in the present study.

Appendix 6.D Relationships with land cover classes

Spearman's rank correlation coeffitients between proability of presence of vectors and percentage cover of land cover types are given in Tables 6.D.1 and 6.D.2.

Table 6.D.1: Spearman's rank correlation coefficients between percentage cover of land cover classes and the maximum a posteriori estimate of probability of presence for each vector (part a).

	Ae. cinereus	*An. atroparvus*	*An. claviger*	*Cq. richiardii*	*Cx. pipiens* s.l.	*Cs. annulata*	*Cs. morsitans*
Broadleaved woodland	0.12	-0.13	0.05	0.03	-0.01	0.03	0.17
Coniferous woodland	0.12	-0.02	0.10	0.12	0.10	0.10	0.16
Arable & horticulture	-0.08	0.13	-0.05	-0.10	-0.18	-0.12	-0.20
Improved Grassland	-0.04	-0.20	-0.11	-0.07	-0.10	-0.11	-0.08
Rough Grassland	0.00	-0.08	-0.04	-0.03	-0.06	-0.04	-0.01
Neutral Grassland	0.06	0.00	0.02	0.07	0.01	0.03	0.03
Calcareous Grassland	-0.01	-0.05	0.00	-0.06	-0.00	-0.01	-0.01
Acid Grassland	-0.08	-0.08	0.01	-0.08	0.01	-0.05	-0.02
Fen, Marsh & Swamp	0.04	0.01	0.04	0.04	0.04	0.04	0.04
Heather	0.03	-0.04	0.03	0.05	0.05	0.04	0.07
Heath Grassland	-0.01	-0.07	0.01	0.00	0.03	-0.00	0.05
Bog	-0.01	-0.03	-0.01	-0.01	0.01	-0.02	-0.01
Montane Habitats	-0.01	0.02	-0.00	-0.00	-0.00	-0.01	-0.01
Inland Rock	0.00	0.02	0.04	0.00	0.04	0.05	0.06
Saltwater	0.06	0.07	0.07	0.07	0.08	0.07	0.07
Freshwater	0.07	0.07	0.08	0.09	0.10	0.10	0.10
Supra-littoral Rock	0.02	0.02	0.02	0.02	0.02	0.01	0.02
Supra-littoral Sediment	0.06	0.06	0.06	0.05	0.06	0.06	0.06
Littoral Rock	0.03	0.03	0.03	0.03	0.03	0.03	0.03
Littoral Sediment	0.05	0.09	0.07	0.06	0.08	0.08	0.07
Saltmarsh	0.05	0.09	0.08	0.06	0.08	0.08	0.07
Urban	-0.01	0.09	0.02	0.03	0.12	0.11	0.06
Suburban	-0.08	0.08	-0.05	-0.03	0.11	0.05	0.08
Coastal & Floodplain Grazing Marsh	0.10	0.16	0.15	0.19	0.16	0.18	0.10
Coastal Sand Dunes	0.03	0.04	0.03	0.02	0.04	0.04	0.03
Coastal Vegetated Shingle	0.02	0.02	0.03	0.02	0.04	0.04	0.03

Table 6.D.2: Spearman's rank correlation coefficients between percentage cover of land cover classes and the maximum a posteriori estimate of probability of presence for each vector (part b).

	Oc. cantans/annulipes	*Oc. detritus*	*Oc. geniculatus*	*Oc. punctor*	*Oc. rusticus*	WNV community
Broadleaved woodland	0.16	0.00	0.19	0.15	0.15	0.13
Coniferous woodland	0.15	0.06	0.13	0.16	0.16	0.15
Arable & horticulture	-0.16	0.10	-0.15	-0.12	-0.26	-0.21
Improved Grassland	-0.15	-0.31	-0.22	-0.21	-0.11	-0.07
Rough Grassland	-0.04	-0.08	-0.08	-0.04	-0.02	-0.01
Neutral Grassland	0.02	-0.01	-0.02	0.01	0.04	0.03
Calcareous Grassland	-0.02	-0.08	-0.01	0.01	-0.04	-0.01
Acid Grassland	-0.07	-0.18	-0.05	0.00	-0.07	-0.03
Fen, Marsh & Swamp	0.04	0.03	0.01	0.04	0.03	0.04
Heather	0.03	-0.05	-0.01	0.09	0.05	0.06
Heath Grassland	-0.00	-0.08	-0.02	0.06	0.02	0.03
Bog	-0.01	-0.11	0.01	0.04	-0.01	-0.01
Montane Habitats	-0.03	-0.05	-0.01	0.01	-0.02	-0.01
Inland Rock	0.04	0.02	0.04	0.05	0.04	0.05
Saltwater	0.07	0.06	0.08	0.06	0.08	0.07
Freshwater	0.10	0.11	0.09	0.09	0.11	0.11
Supra-littoral Rock	0.02	0.02	0.02	0.02	0.02	0.02
Supra-littoral Sediment	0.07	0.06	0.06	0.05	0.07	0.06
Littoral Rock	0.03	0.04	0.03	0.03	0.03	0.03
Littoral Sediment	0.08	0.08	0.08	0.06	0.09	0.08
Saltmarsh	0.08	0.07	0.08	0.06	0.09	0.08
Urban	0.09	0.19	0.19	0.06	0.15	0.07
Suburban	0.11	0.24	0.19	0.03	0.21	0.07
Coastal & Floodplain Grazing Marsh	0.09	0.06	-0.05	0.07	0.09	0.15
Coastal Sand Dunes	0.04	0.04	0.03	0.03	0.04	0.03
Coastal Vegetated Shingle	0.03	0.04	0.04	0.01	0.03	0.03

Table 6.D.3: Regression coefficients (β), χ^2 test statistics and p values of likelihood ratio tests on univariate generalized linear models. All tests performed on one degree of freedom.

Vector	Land cover type	β	χ^2	p <
Aedes cinereus	Broadleaved woodland	-0.302	0	0.985
Aedes cinereus	Coniferous woodland	0.674	0.004	0.949
Aedes cinereus	Coastal & Floodplain Grazing Marsh	-0.368	0	0.997
Anopheles atroparvus	Arable & horticulture	-0.206	5.575	0.018
Anopheles atroparvus	Coastal & Floodplain Grazing Marsh	0.877	25.463	0.001
Anopheles claviger	Coniferous woodland	0.197	0.109	0.741
Anopheles claviger	Coastal & Floodplain Grazing Marsh	1.009	11.315	0.001
Coquillettidia richiardii	Coniferous woodland	0.48	0.031	0.86
Coquillettidia richiardii	Coastal & Floodplain Grazing Marsh	0.624	0.118	0.731
Culex pipiens s.l.	Coniferous woodland	0.477	1.246	0.264
Culex pipiens s.l.	Urban	0.211	1.331	0.249
Culex pipiens s.l.	Suburban	-0.067	1.486	0.223
Culex pipiens s.l.	Coastal & Floodplain Grazing Marsh	0.847	33.681	0.001
Culiseta annulata	Coniferous woodland	0.349	1.734	0.188
Culiseta annulata	Freshwater	1.038	2.784	0.095
Culiseta annulata	Urban	0.155	0.862	0.353
Culiseta annulata	Coastal & Floodplain Grazing Marsh	0.937	53.571	0.001
Culiseta morsitans	Broadleaved woodland	0.284	0.01	0.918
Culiseta morsitans	Coniferous woodland	0.404	0.011	0.916
Culiseta morsitans	Freshwater	0.939	0.051	0.821
Culiseta morsitans	Coastal & Floodplain Grazing Marsh	0.146	0.001	0.973
Ochlerotatus cantans/annulipes	Broadleaved woodland	0.412	0.02	0.887
Ochlerotatus cantans/annulipes	Coniferous woodland	0.551	0.027	0.87
Ochlerotatus cantans/annulipes	Suburban	-0.107	0.001	0.975
Ochlerotatus detritus	Freshwater	0.423	0	0.993
Ochlerotatus detritus	Urban	0.416	0	0.984
Ochlerotatus detritus	Suburban	0.231	0	0.985
Ochlerotatus geniculatus	Broadleaved woodland	0.41	0.024	0.876
Ochlerotatus geniculatus	Coniferous woodland	0.345	0.007	0.934
Ochlerotatus geniculatus	Urban	0.421	0.012	0.914
Ochlerotatus geniculatus	Suburban	0.121	0.002	0.961
Ochlerotatus punctor	Broadleaved woodland	0.386	1.111	0.292
Ochlerotatus punctor	Coniferous woodland	1.054	1.901	0.168
Ochlerotatus rusticus	Broadleaved woodland	0.222	0.005	0.944
Ochlerotatus rusticus	Coniferous woodland	0.423	0.01	0.92
Ochlerotatus rusticus	Freshwater	0.741	0.016	0.898
Ochlerotatus rusticus	Urban	0.199	0.001	0.97
Ochlerotatus rusticus	Suburban	0.105	0.001	0.971
WNV community	Broadleaved woodland	0.003	0.026	0.871
WNV community	Coniferous woodland	0.478	1.302	0.254
WNV community	Freshwater	1.265	9.005	0.003
WNV community	Coastal & Floodplain Grazing Marsh	0.842	14.808	0.001

Appendix 6.E Estimating prevalence

Table 6.E.1 gives the expert opinion prevalence estimates for each species modelled. The instructions given to the expert to estimate prevalences are provided below.

Table 6.E.1: Expert-opinion prevalence estimates and corresponding parameters of the beta distribution calculated using the discrete-habitat prevalence elicitation method.

	mode	95% credible interval	α	β
Aedes cinereus	0.0002	(0.0001, 0.0015)	3.91	14074.49
Anopheles atroparvus	0.0052	(0.0022, 0.0138)	4.56	687.15
Anopheles claviger	0.0065	(0.0039, 0.0134)	14.77	2092.29
Coquillettidia richiardii	0.0119	(0.0076, 0.0227)	19.07	1497.95
Culex pipiens s.l.	0.0637	(0.0441, 0.0931)	25.08	354.7
Culiseta annulata	0.066	(0.0439, 0.0933)	29.88	409.54
Culiseta morsitans	0.0069	(0.0042, 0.015)	15.33	2069.04
Ochlerotatus cantans/annulipes	0.0048	(0.002, 0.0176)	5.39	917.45
Ochlerotatus detritus	0.0009	(0.0003, 0.0042)	2.21	1399.72
Ochlerotatus geniculatus	0.0065	(0.0025, 0.018)	4.24	499.9
Ochlerotatus punctor	0.0045	(0.0023, 0.0099)	6.7	1272.22
Ochlerotatus rusticus	0.0068	(0.0036, 0.0142)	7.64	969.78

Prevalence

In order to produce maps of the probability of presence of mosquito species in the UK using information only on where they are found, we need an estimate of the total proportion of grid cells in which larvae are present. In order to estimate this overall prevalence of the species, we are asking for expert opinion estimates of prevalence in each of a number of land cover types. To ensure that the maps are interpretable, it is important to be clear about what these estimates represent.

When filling in the table in the file *prevalence_table.xlsx* please bear the following scenario in mind:

> A large number of 190m by 190m square plots are randomly demarcated in **England**. If you were to thoroughly survey each of the plots with land cover type x in a single day at the **optimum time of year** for the species, in what proportion of these plots would you expect to find **LARVAE** of mosquito species y?

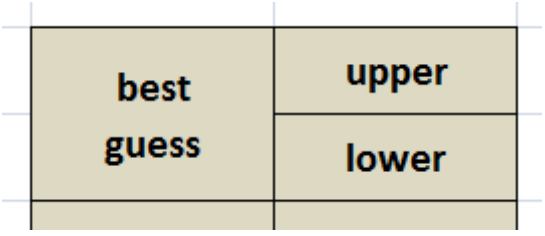
Land cover types

The land cover types provided are all mutually exclusive, with 'coastal and floodplain grazing marsh', 'coastal sand dunes' and 'coastal vegetated shingle' superseding other habitat types. For example the 'rough grassland' category does not include rough grassland in coastal and floodplain grazing marshes. Most of the land cover types will be self-explanatory but technical definitions are provided in the pdf file *habitat_descriptions.pdf*

Whilst each plot has a distinct land cover type, there may be other factors which influence the suitability for mosquito larvae, for example deciduous woodland on steep slopes may be less suitable than deciduous woodland on flat ground. There may also be suitable microhabitats, such as water containers in an otherwise dry arable habitat. Try to use your field experience to incorporate these differences into your estimates.

Filling in the table

For each habitat/species combination there is a box with space for your *best guess* at this proportion and two boxes for *upper and lower bounds* on this estimate which allow your to express your uncertainty in this estimate. The upper and lower bounds should represent values between which you are 95% confident the true proportion lies. Be as precise or imprecise as you wish.



For example, if you think that the most likely proportion of coniferous woodland plots which contain *Cs. annulata* is 0.1, and you're 95% confident the proportion wouldn't be above 0.3 or below 0.01, then fill out the boxes as follows:

	Culiseta annulata	
coniferous woodland	0.1	0.3
		0.01

For many species/habitat combinations the expected proportion may be very low (say less than 1 in 1000; 0.001). In these cases, please enter 0 for the best guess and leave the other two cells blank.

If you do not feel confident making estimates for certain species, even with large confidence intervals, please leave those columns blank.

Chapter 7

General discussion

The following general discussion recapitulates the content and principal findings of this thesis, discusses some relevant issues and draws overall conclusions. The structure of the discussion reflects the two principal areas covered: species distribution modelling (SDM) and mosquito-borne disease risk in the UK.

Recapitulation

Species distribution modelling

Chapter 3 developed a new species interaction distribution model (SIDM) using Bayesian multivariate binomial regression. This model allows the user to investigate correlations between the distributions of multiple species whilst removing the confounding effects of environmental predictors.

Chapter 4 illustrated the distinct problems of calibration bias and contamination of controls when predicting the probability of species presence from presence-background data. Methods were developed to produce accurate estimates of a species' prevalence using the opinions of field experts and to formulate these estimates into valid probability distributions. Two methods were then proposed to incorporate this information into distribution models: a very simple calibration-corrected naïve approach for point estimates of prevalence which, whilst theoretically still biased, produces accurate predictions of probability of presence under reasonable conditions, and a more involved Markov chain Monte Carlo (MCMC) approach which could be used to adapt any parametric SDM for use with presence-background data.

Chapter 5 introduced GRaF, a Bayesian SDM approach using Gaussian random fields (GRFs) to fit complex, but smooth (and hopefully biologically plausible) responses to environmental covariates. GRaF provides methods to account for uncertainty in occurrence records, incorporate knowledge of the species' ecology and automatically produce estimates of prediction uncertainty.

UK mosquito distributions

Chapter 2 reported the establishment of substantial populations of the major European West Nile virus (WNV) vector mosquito *Culex modestus* at wetland sites in southern England. Also present at these sites are the enzootic WNV maintenance vector *Cx. pipiens*, another potential bridge vector *Culiseta annulata* and the former European malaria vector *Anopheles atroparvus*.

Chapter 3 applied the novel SIDM to identify environmental conditions and biotic interactions which drive the distribution of this potential WNV vector community in UK wetlands. Environmental conditions such as water depth are the most important drivers, though two predators of mosquito larvae also appear to influence the spatial distribution of this community.

Chapter 6 applied GRaF to a comprehensive database of mosquito occurrence records to produce the first high-resolution distribution maps for UK mosquitoes. Prevalence estimates were elicited from expert opinion and incorporated into these models using methods developed earlier in the thesis, so that these maps provide robust predictions of probability of presence. These predictions were combined to map the probability of presence of potential WNV vector communities in the UK. A number of wetland areas were highlighted in which WNV could probably be maintained in an enzootic cycle and transmitted to humans.

Species distribution modelling

Black-box and open-box models

There is an apparent distinction within species distribution modelling between *correlative* and *process-based* approaches (Elith & Leathwick, 2009). Models in which ecological processes are explicitly defined *a priori* are termed process-based, whereas those which fit more vague, non-deterministic relationships are considered to be correlative (Dormann *et al.*, 2012). Correlative models have been shown to have very good predictive power (Elith *et al.*, 2006), whereas process-based approaches afford more powerful and generalisable explanations of the species' underlying ecology (Kearney & Porter, 2009). However Dormann *et al.* (2012) note that for many modelling approaches this apparent dichotomy is not so clear-cut, with different approaches better described by a 'process-correlation continuum'.

Even among the more correlative models, there exists a continuum between those with stronger predictive and explanatory abilities. Comparatively 'black-box' approaches such as random forests and artificial neural networks produce models with high predictive power, but which resist direct ecological interpretation (Paruelo & Tomasel, 1997; Cutler *et al.*, 2007). By contrast, 'open-box' approaches such as generalised linear models allow clear interpretation of fitted parameters and can be interrogated using a range of hypothesis testing procedures (Zuur *et al.*, 2009), though they perform less well at predictive tasks (Elith *et al.*, 2006).

This thesis has developed and applied correlative SDMs at both ends of this continuum. The community model used in Chapter 3 was relatively open-box, with parameters corresponding directly to biotic and abiotic relationships (though not the mechanisms underlying them). This approach had good explanatory power: the strength and evidentiary support of these relationships could easily be examined. By contrast, the GRaF approach developed in Chapter 5 and applied in Chapter 6 is comparatively black-box. Whilst GRaF models can to some extent be visualised and compared, it is more difficult to understand the ecological implications of the fitted

model.

Whilst the clearly-defined model structure enforced by parametric approaches make them easy to interpret, these interpretations are not necessarily accurate. If the analytical model is a poor approximation of of the 'true' model, any inferences from it will be biased (Anscombe, 1973; Zuur *et al.*, 2009). By enforcing less strict structural assumptions, approaches such as GRaF, boosted regression trees and artificial neural networks may produce more accurate representations of the underlying ecological phenomena. Improving our ability to extract generalisable ecological knowledge from these models would help to bridge the dichotomy between black-box and open-box correlative SDMs (Olden & Jackson, 2002).

Opening up GRaF

GRaF's coherent Bayesian statistical approach allows for a number of extensions which could improve interpretability of the fitted models. Additive GRF models (Duvenaud *et al.*, 2011) would allow an exploration of the degree of complexity present in the species' response to its environment. Comparing additive GRFs with different levels of interaction complexity would allow us to examine the dimensionality of a species' response to its environment and to identify which covariate interactions are most important. GRaF models constructed as an additive combination of (optionally non-additive) sub-models could be examined using functional ANOVA procedures (Kaufman & Sain, 2010). This would allow a robust statistical assessment of the relative importance of different groups of explanantory variables, such as climatic conditions, anthropogenic impacts or spatio-temporal effects. Multiple-reponse GRFs (Osborne *et al.*, 2008; Alvarez & Lawrence, 2009; Alvarez, 2011) would enable us to construct GRaF SIDMs. Inter-species correlation coefficients could then be parameterised (as in Chapter 3) in conjunction with GRaF models of each species' environmental niche.

The current implementation of GRaF would allow fitted process-based SDMs to be incorporated as existing ecological knowledge, via the models' mean function. This

could be used to increase the predictive power of the model by incorporating non-environmental processes into predictions However this is very much a one-way process, GRaF provides no additional information about the modelled processes. More comprehensive integration of process-based approaches with GRaF would require simultaneous parameterisation of the two components. Depending on the processes in question this may be difficult to implement, but such a model could improve on both the predictive and explanatory power of current approaches.

Interpretating presence-background predictions

For many applications of presence-background SDM, predictions of habitat suitability are sufficient. Such models enable identification of the major ecological correlates of distributions and identification of sites in which the species is *most likely* to be present. But where the aim is to predict absolute probability of presence from presence-background data, the naïve approaches which are currently in wide use are clearly inadequate. The correct interpretation of predictions from un-calibrated presence-background models as a species-specific index of habitat suitability is clearly stated in one of the most highly cited recent papers on SDM (Elith *et al.*, 2006). Unfortunately, numerous applications of presence-background SDM continue to misinterpret the resulting distribution maps as representing probability of presence. Yackulic *et al.* (2012) found that a 54% of uncorrected MaxEnt distribution maps were incorrectly labelled as probability of presence.

Subjective probability

Interestingly, Elith *et al.* (2011) suggested that uncorrected predictions from presence-background models *could* be interpreted as probability of presence, subject to a liberal interpretation of the sampling conditions (such as detection probability and spatial and temporal sampling scale). This perspective shares similarities with the subjectivist Bayesian interpretation that any probability can be considered valid since it constitutes a subjective personal belief (De Finetti, 1974). The validity of subjec-

tivist and objectivist approaches to probability theory are a subject of continuing philosophical debate. However, the theoretical validity of such an interpretation of presence-background predictions does not imply that it is *useful*. To be interpreted and re-used, predicted probabilities must be accompanied by clearly-stated sampling conditions. In Chapter 6 we achieve this by clearly stating the conditions of probability of presence when eliciting an expert-opinion estimate of each species' prevalence (see Appendix 6.E). Because this interpretation was common across all species, the resulting probabilities could be re-used when predicting the probability of presence of WNV vector communities.

Mosquito-borne disease risk to the UK

This thesis has focussed exclusively on mosquito species already present in the UK and the risks posed by pathogens they could transmit. Here, the overall risk posed by these pathogens is discussed, followed by the risks from exotic vectors.

Risk posed by native vectors

Malaria

Over the last three decades, the annual number of imported cases of *Plasmodium vivax* malaria has decreased by 70%, from around 1,000 to 300 cases (Smith *et al.*, 2008). During the same time period however, importation of the more deadly *P. falciparum* increased 31% to approximately 1,700 cases per annum. The former European malaria vector *Anopheles atroparvus* is known to be a competent vector of *P. vivax* and in 1917 was responsible for a large outbreak of the disease on the Isle of Sheppey, Kent, after infected soldiers were stationed there (Shute, 1945). The competency of *An. atroparvus* for *P. falciparum* is less clear; Shute (1940) found specimens collected in the North Kent Marshes to be susceptible to strains of the parasite originating in temperate regions, but not those from the tropics. It seems however that the risk posed to the UK by either species of malaria is low, due to the highly restricted distribution of *An. atroparvus* (Chapter 6) and its apparent preference for feeding on livestock rather than humans (Marshall, 1938). In the few areas where human biting by *An. atroparvus* where transmission might be possible, it seems likely that introduced malaria could be rapidly eliminated through treatment and vector control (Hutchinson, 2004).

Arboviruses

The relatively widespread distribution of the known maintenance vector *Cx. pipiens* and of susceptible bird species suggests that enzootic cycles of WNV and ecologically similar pathogens could be maintained in much of the UK, provided a sufficient abundance of the hosts and vectors. Transmission of such diseases to humans and livestock

would be limited to those wetland areas where suitable bridge vectors, such as *Cx. modestus* occur. Since wetlands tend to host the highest abundances of mosquitoes in the UK, as well as large avian populations, these areas also seem the most likely to maintain enzootic cycles. European wetlands also host large numbers of migratory birds which could import viruses from areas where arboviruses are established (Guillemain *et al.*, 2009).

However the comparatively low population densities in European wetlands limits the potential for spread to humans. In those UK wetlands where all the necessary ingredients for WNV transmission coincide, an outbreak of similar severity of recent outbreaks in southern Europe seems unlikely, due to the comparatively short adult biting season and low abundances of mosquitoes in the UK.

Exotic vectors

The establishment of *Cx. modestus* in the UK clearly increases the risk of human WNV cases, at least in those areas where it is present. A number of other vector mosquitoes have the potential to become established in the UK.

The Asian tiger mosquito

One of the most concerning species is *Aedes albopictus*, the 'Asian tiger mosquito', an efficient vector of dengue and chikungunya fever viruses. *Ae. albopictus* originates in Southeast Asia but is now also established in North and South America, Africa, Australia and Europe. Twenty European countries have reported the presence of this species (Medlock *et al.*, 2012). The rapid global spread of *Ae. albopictus* - greatly aided by the international trade in used tyres (Hawley *et al.*, 1987) - has earned it a reputation as the most invasive mosquito species in the world. Unlike most UK mosquito species, *Ae. albopictus* tends to be prevalent in urban areas, breeding in the small containers of standing water which tend to be abundant around human habitation. As a result it comes into regular contact with human populations, which it bites aggressively (Becker *et al.*, 2010). *Ae. albopictus* is now considered to be the

major nuisance biting mosquito in Italy (Scholte *et al.*, 2007) and in parts of Spain and France, where the species has only recently arrived (Aranda *et al.*, 2006; Vazeille *et al.*, 2008). Where *Ae. albopictus* has become established in Europe, outbreaks of dengue and chikungunya fevers have swiftly followed (Rezza *et al.*, 2007; La Ruche *et al.*, 2010; Grandadam *et al.*, 2011).

Modelling studies have suggested that parts of southern England have climatic conditions suitable for *Ae. albopictus* to persist (Medlock *et al.*, 2006; Benedict *et al.*, 2007; Caminade *et al.*, 2012). However the UK is at the very edge of these predicted ranges and the risk of the species becoming established is therefore uncertain.

The yellow fever mosquito

Ae. aegypti, the most important global vector of dengue, yellow and chikungunya fever viruses, is currently established around the Black sea in Russia, Georgia and Abkhazia (Medlock *et al.*, 2012). The species was previously established in Western Europe and sightings reported as far north as Brest, though it seems unlikely that *Ae. aegypti* could become established in the UK under current environmental conditions (Reiter (2010), despite a dubious historical report from Epping Forest (MacGregor, 1919)).

Other species

A number of other exotic and invasive aedine mosquitoes have been introduced to Europe and appear to be spreading. These include *Ae. japonicus*, *Ae. koreicus* and *Ae. atropalpus*, though none of these has been proven to be a vector in the field (Medlock *et al.*, 2012).

General conclusions

GRaF, the distribution modelling approach introduced in this thesis provides a number of advances on existing correlative species distribution models (SDMs). By eliciting expert opinion estimates, we can correct GRaF and other SDMs to produce correctly-calibrated maps of probability of presence, using abundant presence-only data. Extensions to GRaF could increase the amount of ecological interpretability of these models.

West Nile virus appears to be the greatest public health risk posed to the UK by native mosquito species. The establishment of *Culex modestus* in wetlands in southern England substantially increases this risk. Experimental evaluation of the impact of ditch shrimp (genus *Palaemonetes*) and fish on potential WNV vector communities in UK wetlands may allow for sustainable mosquito control by promoting these predators.

Based on the limited available information, the risk of WNV transmission appears to be focussed in wetland areas. This thesis has provided a quantitative prediction of areas where transmission of the pathogen to humans could occur. Organized, broad-scale surveys are required to increase our knowledge of the distributions of *Cx. modestus* and other potential vector mosquitoes. A comparison of the distribution of potential WNV vector communities with distributions of avian and human hosts and potential routes for disease introduction is needed to more fully map risk of the disease.

References

Alvarez, M. (2011) *Convolved Gaussian process priors for multivariate regression with applications to dynamical systems*. Ph.D. thesis, University of Manchester.

Alvarez, M.A. & Lawrence, N.D.N. (2009) Sparse convolved multiple output gaussian processes. *arXiv preprint arXiv:09115107*, pp. 1–33.

Anscombe, F. (1973) Graphs in statistical analysis. *The American Statistician*, **27**, 17–21.

Aranda, C., Eritja, R. & Roiz, D. (2006) First record and establishment of the mosquito Aedes albopictus in Spain. *Medical and Veterinary Entomology*, **20**, 150–2.

Becker, N., Petric, D., Zgomba, M., Boase, C., Madon, M., Dahl, C. & Kaiser, A. (2010) *Mosquitoes and Their Control*. Springer Verlag, Berlin, second edition.

Benedict, M.Q., Levine, R.S., Hawley, W.a. & Lounibos, L.P. (2007) Spread of the tiger: global risk of invasion by the mosquito Aedes albopictus. *Vector-Borne and Zoonotic Diseases*, **7**, 76–85.

Caminade, C., Medlock, J.M., Ducheyne, E., McIntyre, K.M., Leach, S., Baylis, M. & Morse, A.P. (2012) Suitability of European climate for the Asian tiger mosquito Aedes albopictus: recent trends and future scenarios. *Journal of the Royal Society, Interface*, **9**, 2708–17.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J. & Lawler, J.J. (2007) Random forests for classification in ecology. *Ecology*, **88**, 2783–92.

De Finetti, B. (1974) *Theory of probability: a critical introductory treatment*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley.

Dormann, C.F., Schymanski, S.J., Cabral, J., Chuine, I., Graham, C.H., Hartig, F., Kearney, M., Morin, X., Römermann, C., Schröder, B. & Singer, A. (2012) Correlation and process in species distribution models: bridging a dichotomy. *Journal of Biogeography*, **39**, 2119–2131.

Duvenaud, D., Nickisch, H. & Rasmussen, C. (2011) Additive Gaussian processes. *arXiv preprint arXiv:11124394*, pp. 1–9.

Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M.M., Townsend Peterson, A., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, Robert, E., Soberón, J., Williams, S., Wisz, M.S. & Zimmermann, N.E. (2006) Novel methods improve prediction of species distributions from occurrence data. *Ecography*, **29**, 129–151.

Elith, J. & Leathwick, J.R. (2009) Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 677–697.

Elith, J., Phillips, S. & Hastie, T. (2011) A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, **17**, 43–57.

Grandadam, M., Caro, V., Plumet, S., Thiberge, J.M., Souarès, Y., Failloux, A.B., Tolou, H.J., Budelot, M., Cosserat, D., Leparc-Goffart, I. & Desprès, P. (2011) Chikungunya virus, southeastern France. *Emerging infectious diseases*, **17**, 910–3.

Guillemain, M., Hearn, R. & King, R. (2009) Comparing the migration of Eurasian Teal Anas crecca from two main wintering areas of Western Europe: A long-term study from Essex, England, and the Camargue,. *Ringing and Migration*, **24**, 273–276.

Hawley, W.A., Reiter, P., Copeland, R.S., Pumpuni, C.B. & Craig, G.B. (1987) Aedes albopictus in North America: probable introduction in used tires from northern Asia. *Science*, **236**, 1114–6.

Hutchinson, R.A. (2004) *Mosquito Borne Diseases in England: past, present and future risks, with special reference to malaria in the Kent Marshes.* Ph.D. thesis, Durham.

Kaufman, C. & Sain, S. (2010) Bayesian functional ANOVA modeling using Gaussian process prior distributions. *Bayesian Analysis*, **5**, 123–150.

Kearney, M. & Porter, W. (2009) Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology Letters*, **12**, 334–50.

La Ruche, G., Souarès, Y., Armengaud, A., Peloux-Petiot, F., Delaunay, P., Desprès, P., Lenglet, A., Jourdain, F., Leparc-Goffart, I., Charlet, F., Ollier, L., Mantey, K., Mollet, T., Fournier, J.P., Torrents, R., Leitmeyer, K., Hilairet, P., Zeller, H.G., Van Bortel, W., Dejour-Salamanca, D., Grandadam, M. & Gastellu-Etchegorry, M. (2010) First two autochthonous dengue virus infections in metropolitan France, September 2010. *Euro Surveillance*, **15**, 19676.

MacGregor, M.E. (1919) On the Occurrence of Stegomyia fasciata in a Hole in a Beech Tree in Epping Forest. *Bulletin of Entomological Research*, **10**, 91.

Marshall, J.F. (1938) *The British Mosquitoes.* Trustees of the British Museum.

Medlock, J.M., Avenell, D., Barrass, I. & Leach, S. (2006) Analysis of the potential for survival and seasonal activity of Aedes albopictus (Diptera: Culicidae) in the United Kingdom. *Journal of Vector Ecology*, **31**, 292–304.

Medlock, J.M., Hansford, K.M., Schaffner, F., Versteirt, V., Hendrickx, G., Zeller, H.G.H. & Van Bortel, W. (2012) A review of the invasive mosquitoes in Europe: ecology, public health risks, and control options. *Vector-Borne and Zoonotic Diseases*, **12**, 435–47.

Olden, J.D. & Jackson, D.A. (2002) Illuminating the black box: a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling*, **154**, 135–150.

Osborne, M.A., Roberts, S.J., Rogers, A., Ramchurn, S.D. & Jennings, N.R. (2008) Towards Real-Time Information Processing of Sensor Network Data Using Computationally Efficient Multi-output Gaussian Processes. *2008 International Conference on Information Processing in Sensor Networks*, pp. 109–120.

Paruelo, J.M. & Tomasel, F. (1997) Prediction of functional characteristics of ecosystems: a comparison of artificial neural networks and regression models. *Ecological Modelling*, **98**, 173–186.

Reiter, P. (2010) Yellow fever and dengue: a threat to Europe? *Eurosurveillance*, **15**, 19509.

Rezza, G., Nicoletti, L., Angelini, R., Romi, R., Finarelli, A.C., Panning, M., Cordioli, P., Fortuna, C., Boros, S., Magurano, F., Silvi, G., Angelini, P., Dottori, M., Ciufolini, M.G., Majori, G.C. & Cassone, A. (2007) Infection with chikungunya virus in Italy: an outbreak in a temperate region. *Lancet*, **370**, 1840–6.

Scholte, E., Dijkstra, E., Ruijs, H. & Jacobs, F. (2007) The Asian tiger mosquito (Aedes albopictus) in the Netherlands: should we worry? *Proceedings of the Netherlands Entomological Society Meeting*, **18**, 131–136.

Shute, P.G. (1940) Failure to infect english specimens of anopheles maculipennis var. atroparvus with certain strains of plasmodium falciparum of tropical origin. *Journal of Tropical Medicine and Hygiene*, **43**, 175–178.

Shute, P.G. (1945) Malaria in England. *Public Health*, **58**, 62–65.

Smith, A.D., Bradley, D.J., Smith, V., Blaze, M., Behrens, R.H., Chiodini, P.L. & Whitty, C.J.M. (2008) Imported malaria and high risk groups: observational study using UK surveillance data 1987-2006. *British Medical Journal*, **337**, 103–106.

Vazeille, M., Jeannin, C., Martin, E., Schaffner, F. & Failloux, A.B. (2008) Chikungunya: a risk for Mediterranean countries? *Acta Tropica*, **105**, 200–2.

Yackulic, C.B., Chandler, R.B., Zipkin, E.F., Royle, J.A., Nichols, J.D., Campbell Grant, E.H. & Veran, S. (2012) Presence-only modelling using MAXENT: when can we trust the inferences? *Methods in Ecology and Evolution*, **4**, 236–243.

Zuur, A., Ieno, E., Walker, N. & Saveliev, A. (2009) *Mixed effects models and extensions in ecology with R.* Springer Science, New York.