Wetlands (2017) 37:205–219
DOI 10.1007/s13157-016-0859-z

ORIGINAL RESEARCH

# Disaggregating the Patchwork:

## Probabilistic models as tools to predict wetland presence as a continuous gradient

John M. Humphreys[1,2] · James B. Elsner[2] · Thomas H. Jagger[2] · AmirSassan Mahjoor[2]

Received: 4 June 2016 / Accepted: 30 November 2016 / Published online: 21 December 2016

**Abstract** Identification and inventory of wetlands are essential components of natural resource management. To be effective in these endeavors, it is critical that the process used to detect and document wetlands be time efficient, accurate, and repeatable as new environmental information becomes available. Approaches dependent on aerial photographic interpretation of land cover by individual human analysts necessitate hours of assessment, introduce human error, and fail to include the best available soils and hydrologic data. The goal of the current study is to apply hierarchical modeling and Bayesian inference to predict the probability of wetland presence as a continuous gradient with the explicit consideration of spatial structure. The presented spatial statistical model can evaluate 100 km$^2$ at a 50 x 50 meter resolution in approximately 50 minutes while simultaneously incorporating ancillary data and accounting for latent spatial processes. Model results demonstrate an ability to consistently capture wetlands identified through aerial interpretation with greater than 90 % accuracy (scaled Brier Score) and to identify wetland extents, ecotones, and hydrologic connections not identified through use of other modeling and mapping techniques. The provided model is reasonably robust to changes in resolution, areal extents between 100 km$^2$ and 300 km$^2$, and region-specific physical conditions.

✉ John M. Humphreys
john.humphreys@dep.state.fl.us

[1] Florida Department of Environmental Protection,
2600 Blair Stone Road, Tallahassee, Fl 32399, USA

[2] Department of Geography, Florida State University,
Tallahassee, FL, USA

**Keywords** Wetland · Hierarchical model · Bayesian · Landscape · Patch matrix · Spatial correlation · Random-effect · Gradient-based · INLA

## Introduction

Wetlands provide a vast array of ecological services, functions, and values. From flood protection and the maintenance of water quality, to carbon sequestration, the mediation of biogeochemical cycles, and numerous economic, aesthetic, and cultural benefits, wetlands are paramount to human health and well-being (Mitsch and Gossilink 2000, Barbier 2011). Notwithstanding the widely-recognized need to identify and manage wetland resources in a sustainable way (Christensen 1996, Turner 1989), techniques to efficiently inventory and realistically model wetland presence at the landscape-level remain elusive (Finlayson et al. 1999, Rebelo et al. 2009, Meixler and Bain 2010).

The National Wetland Inventory (NWI) produced by the United States Department of Interior Fish and Wildlife Service is the most widely used source of spatial wetland data in the United States and has been utilized to predict impacts from sea-level rise, to undertake wetland restoration planning, and to perform ecological modeling for a variety of wildlife species and habitats (Fish and Wildlife Service 2009). Although the NWI has been applied to a diverse range of studies, its creation and continued development are dependent on the time-consuming aerial interpretation of vegetative cover by individual human analysts (Dahl 2011).

Vegetative land cover can play an important role in wetland identification (Adam et al. 2010); however, vegetative evidence in the absence of other corroborative data may underestimate wetland presence or extent. For example, indicators of hydrology as derived from elevation models may be valuable in determining wetland presence (Lang et al. 2012, Lang et al. 2013) and the ability of soil chemistry and morphology to reflect hydrologic gradients is a key tenant of hydropedology (Pennock et al. 2014). Indeed, the robustness of the hydrology, soils, and vegetation triad of wetland indicators is well-established and even has been legally formalized into the regulatory definition of wetlands as implemented in Section 404 of the Clean Water Act (33 U.S.C. §1251 et seq. 1972). In Florida for example, Chapter 62-340 of the Florida Administrative Code details that delineation of the landward extent of wetlands or other surface waters for regulatory purposes requires consideration of a site's hydrology, soils, and vegetation. The essence of the rule is that if evidence for "two out of three" of these indicators is present at a location, then that area may be designated as a wetland or surface water. Even though elevation and soils data are increasingly available at no cost, reliance on qualitative methods bars the inclusion of such ancillary information from the wetland identification process. Photographic interpretive techniques, like those used in development of the NWI, may not fully exploit the potential time-savings, scientific rigor, and ecological realism available through quantitative landscape ecology and spatial statistics.

Landscapes are spatially heterogeneous. The charge of the landscape ecologist is to explain this heterogeneity in terms of pattern and process with explicit examination of spatial structure, variability, and scale (Wiens 1989). In spite of the need for consideration of spatial structure in ecological modeling (Hoeting 2009), efforts to improve on the NWI have focused largely on aggregation approaches in which NWI data are merged with selected attributes from National Resources Conservation Service (NRCS) soil maps, the National Hydrography Dataset (NHD), or similar sources using attribute look-up tables or "spatial join" procedures in a Geographic Information Systems (GIS) environment (see e.g. Galbraith et al, 2003; Reif et al, 2009; Dvorett et al, 2012). Spatial processes are often excluded in these efforts and the selection criteria used by researchers in choosing what attributes to amalgamate are rarely provided or detailed. In a few instances, investigations have included spatial statistics in a post hoc fashion to help describe aggregated results (see e.g. Mccauley and Jenkins, 2005; Martin et al, 2012); however, the unambiguous consideration of spatial dependence, spatial autocorrelation, or other latent structural processes as potential explanatory variables in predicting wetland presence is uncommon. Beyond neglecting to leverage the explanatory power of spatial processes, the majority of wetland modeling studies generate results under the patch matrix perspective (however, see Murphy et al, 2007). That is, most wetland models reduce real-world environmental gradients to dichotomous units of wetland presence or absence.

Since the publication of *Patches and Structural Components for A Landscape Ecology*, (Forman and Godron, 1981), patch matrix has been the prevailing method used to represent landscape pattern. Despite wide-ranging endorsement and implementation, the patch matrix model's basis in delineation of discrete ecological units or "patches" is problematic under many modeling scenarios. For instance, patch matrix boundaries are often subjectively determined, they may not be relevant to focal ecological processes, and they often fail to represent landscapes in an ecologically realistic context (McGarigal 2005, McGarigal et al. 2009, Evans and Cushman 2009, Lausch et al. 2015). Furthermore, the homogenization or simplification of landscape structure diminishes measurable spatial heterogeneity and thereby decreases the total amount of information available to unravel the pattern and process relationship (McGarigal 2005). In contrast to the patch matrix model's categorization of landscapes as patchworks of discrete ecological units, the gradient-based approach aims to quantify landscape heterogeneity through analysis of continuous value ranges.

Continuous values better represent environmental variability precisely because they capture the trends, gradations, and transitions found within and among landscape components (McGarigal 2005). In the case of wetlands, the gradient-based perspective allows for wetlands to be modeled and represented as integrated components of the larger hydrologic system, complete with ecotones, riparian zones, transitional areas, and hydrologic connections (Murphy et al. 2007). As part of a larger theoretical framework, movement from the patch matrix model to a gradient-based methodology tracks the historical trajectory of landscape ecology. That is, landscape ecology has matured from a once descriptive science, concerned predominantly with the taxonomy of homogeneous types or units, to one focused on the reciprocity of pattern and process as observed in multiple dimensions and restrained by scale (Wiens 1989, Wu and Loucks 1995, Wu and Hobbs 2002, 2007).

Capable of accommodating both fixed and random effects, Bayesian probability models may offer resolution to many of the difficulties encountered when predicting wetland presence. The tiered configuration (hierarchy) of the probability models allow for incorporation of "known" environmental variables as well as the "unknown" effects associated with latent processes like spatial correlation. Within the hierarchical model framework, latent structural processes can be assimilated into models via a random-effect term that serves to quantify the uncertainty remaining after accounting for the effects of fixed environmental covariates (Elsner et al. 2016). Until recently, fitting of Bayesian hierarchical models has been restricted to computationally demanding Markov chain Monte Carlo (MCMC) simulation; however, integrated nested Laplace approximation (INLA) uses an approximation for inference and therefore provides a newly accessible and fast alternative to MCMC (Rue et al. 2009).

The goal of the current study is to employ techniques from spatial statistics to predict the probability of wetland presence as a continuous gradient and with explicit consideration of spatial structure. As a first step to accomplishing this task, explanatory variables linked to soils, hydrology, and vegetation are extracted from freely available government datasets. Following verification of each individual variable's significance in explaining the presence of wetlands, the variables are fitted to a hierarchical model. In addition to the soil, hydrologic, and vegetative fixed effects, a Gaussian random-effect is stacked into the model to incorporate latent spatial structure as a stochastic process. Finally, model robustness is evaluated through examination of its sensitivity to changes in scale, extent, and geographic location.

## Methods

### Domain and Data

Model fitting is conducted for a primary domain in northwest Florida. Model robustness is then assessed by applying the best performing model from the primary domain to several other resolutions, areal extents, and geographic locations.

The primary domain has an extent of 100 km$^2$ and is centered on Wakulla Springs State Park, approximately 23 km south of Tallahassee, Florida [Fig. 1]. The area is bounded by a 10 km by 10 km square with a northwest corner located at 30.23° N Latitude and $-84.25°$ W Longitude. The domain occurs in the Southeast Coastal Plain and is within the Gulf Coastal Lowlands physiographic province. Elevation over the area is mostly uniform between three and nine meters above sea level with infrequent increases to a maximum of approximately fourteen meters. The Wakulla River and its associated floodplain transect the study area from the northwest to the southeast corners. Dominant land cover over the domain includes natural upland communities, natural wetland communities, silvaculture, agricultural, and developed areas (commercial and residential).

The second and third domains overlap that of the primary study area, but encompass larger areal extents. These domains double and then triple the area of the primary domain to 200 km$^2$ and 300 km$^2$ respectively and are used to evaluate model sensitivity to changes in areal extent. The fourth domain [Fig. 2] is also located within the Gulf Coastal Lowlands physiographic province in northwest Florida, but does not overlap the primary study area. The fifth and final domain [Fig. 3] straddles the Peace River in peninsular Florida about 80 km due east of the City of Bradenton. Both the fourth and fifth domains exhibit an areal extent of approximately 100 km$^2$.

Analysis and modeling are performed using the open-source R language for statistical computing (R Core Team 2016) with freely-available government data. Incorporated data includes the 1/3 arc sec digital elevation model from the United States Geological Service (USGS) [http://ned.usgs.gov/], surface reflectance from the National Aeronautics and Space Administration and USGS Landsat 8 collaboration [http://landsat.usgs.gov/landsat8.php], land use data from the Florida Department of Environmental Protection [http://www.dep.state.fl.us/gis/datadir.htm], and Soil Survey Geographic Database soils data (SSURGO) from the NRCS [http://websoilsurvey.sc.egov.usda.gov/App/HomePage.htm]. Copies of the R code and all data used in the study are available at github.com/JMHumphreys/WetlandPatchwork.

### Preliminary Analysis

To begin, soils data representing the primary domain are spatially subset based on their coincidence with natural upland and wetland land uses as identified by overlaying land use data. In conjunction with the sub-setting of soils data, a binary value (1, 0) designating each soil mapping unit's (SMU) geographic association with a wetland (1) or upland (0) is added to the soils spatial object. In addition to the newly added vector specifying each SMU's affiliation to a wetland or upland land use, the full complement of the dataset's native attributes are maintained. To reveal underlying data structure, native soil attributes are then transformed using principal component and correspondence analyses.

Continuous numeric attributes from the soils spatial object, such as those estimating depth to water table and water storage, are explored using a Principal Components Analysis (PCA). In a similar fashion, categorical and nominal soil attributes (e.g. soil taxa, drainage classification, etc) are evaluated through Multiple Correspondence Analysis (MCA). Both the PCA and MCA are performed using the **FactoMineR** package (Lê et al. 2008). Next, the decomposed variables from the PCA and MCA are assessed for statistical significance by regressing each onto the binomial land use vector that was generated during the initial sub-setting of soils data. The regression is carried-out using a non-spatial model as implemented using the **r-INLA** package (Rue et al. 2009). Those decomposed soil variables with significant non-zero model coefficients as determined by the credible interval are retained for further investigation.

To supplement the decomposed soil variables, estimates for Available Water Capacity (AWC) and Organic Mass (OM) are calculated using the **soilDB** package (Beaudette et al. 2016). This is done by aggregating attributes by the horizon-level and then by the soil profile before performing

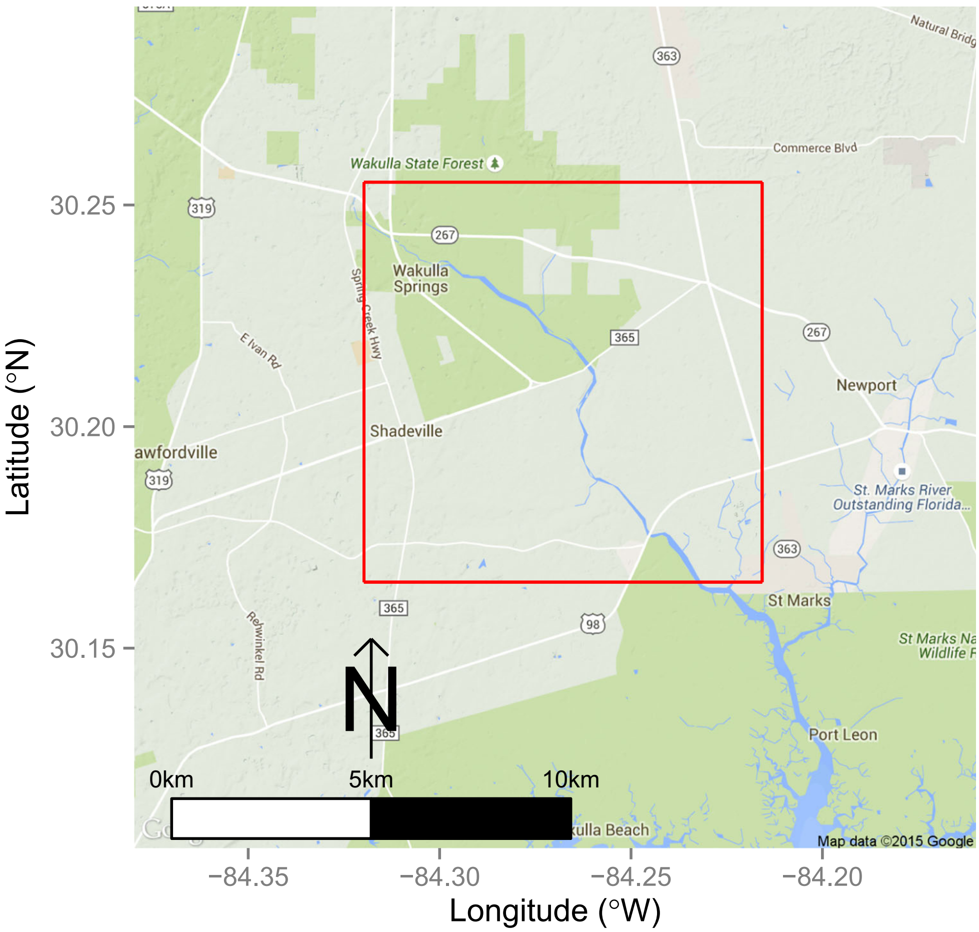

**Fig. 1** Primary Domain (Domain 1). The 100 km$^2$ domain is located in Wakulla County, Florida. The background map is produced using functions in the ggmap package (Kahle and Wickham 2013)

a weighted average across each SMU based on the percent composition attribute. AWC is expressed as a volume fraction and, after adjusting for salinity and inclusive fragments, represents the difference in tension between water contents at field capacity (typically, one-tenth to one-third bar) and fifteen bars (Veihmeyer and Hendrickson 1927). OM describes the amount of decomposed organic matter present as a weighted percentage of soil material. As with the decomposed soils data, the significance of AWC and OM in predicting wetland presence within the domain is verified using non-spatial regression. There is concern in regard to the stability of the model coefficients due to correlation between attributes used in estimating AWC and those decomposed from the original soils dataset; however, these potentially confounding effects are addressed during final model fitting and are discussed below.

Having identified potential variables linked to soils, a Soil Adjusted Vegetation Index (SAVI) is constructed as a proxy for the wetland vegetation found in the domain using surface reflectance from the National Aeronautics and Space Administration and USGS Landsat 8 collaboration (30$m^2$ resolution). The fluctuation of chlorophyll concentration within wetland vegetation is correlated to site-specific hydrology, and because chlorophyll reflects the red and near-infrared spectra disproportionately, remotely sensed reflectance data can be applied to identify possible wetlands (Adam et al. 2010). The SAVI quantifies the ratio of the red and near-infrared bands of surface reflectance by dividing the difference of the two by their sum plus a standard correction factor (Masek et al. 2006). Following calculation of the SAVI, its ability to predict wetland presence is confirmed using non-spatial regression as previously described.

Recognizing that digital elevation data can be used to identify surface flow and wetland hydrology (Moore et al. 1993), two topographic indexes are derived. First, a simple Topographic Position Index (TPI) is calculated as the difference between the value of a target cell and the mean value of its eight neighboring cells. Secondly, a Compound Topographic Index (CTI) is given as the natural log of the upstream contributing area divided by the tangent of the slope for each cell. The explanatory power of both the TPI and CTI are checked through non-spatial regression.

As a final step before model fitting, a spatial adjacency graph is constructed for the study domain using the

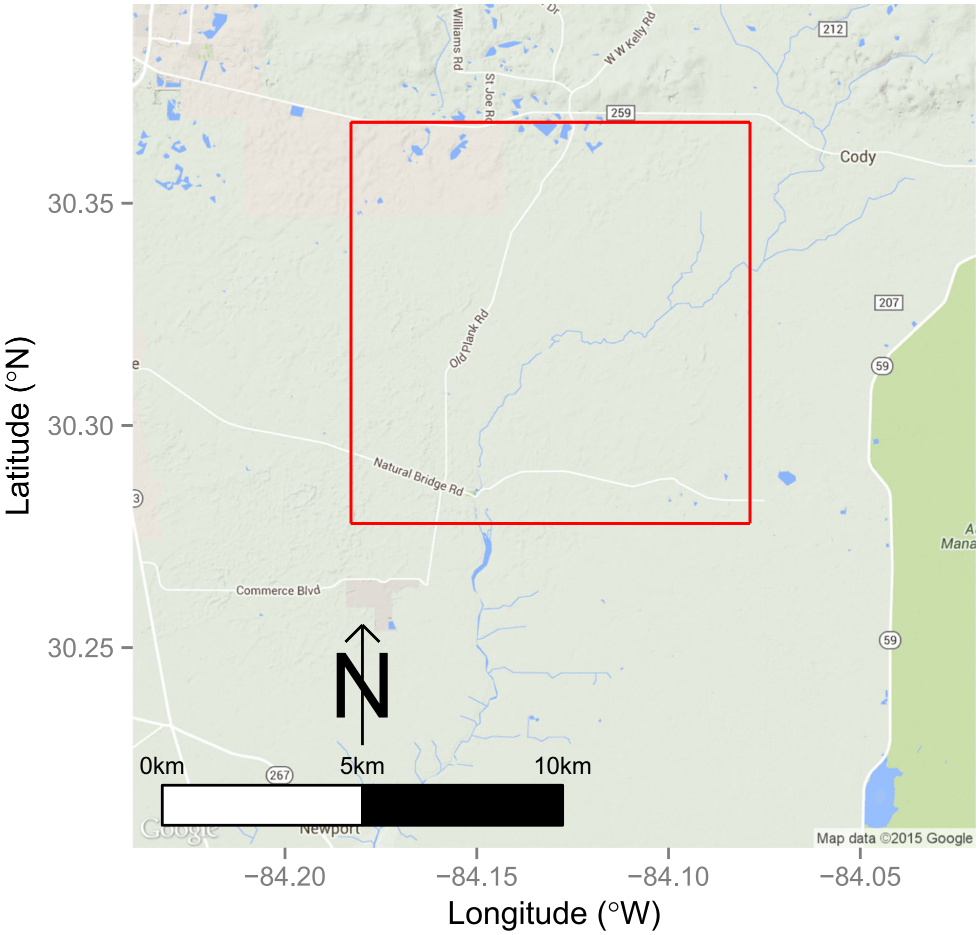

**Fig. 2** Domain 4 located in northwest Florida. The 100 km$^2$ domain is found in Leon County, Florida

**spdep** package (Bivand and Piras 2015). Determination of neighbor contiguity during graph construction is restricted to a minimum of two neighboring cells and a maximum of eight (i.e. a "queen" configuration).

## Model Fitting and Selection

The probability of wetland presence ($p_s$) at cell $s$ is given as

$$\pi(p_s) \sim \text{Binomial}(\mu_p, \eta), \tag{1}$$

where the presence or absence of a wetland at a location (either 0 or 1) is described by a binomial distribution with mean

$$\mu_p = \frac{\exp(\eta)}{1+\exp(\eta)}, \tag{2}$$

and a mean probability ($\mu_p$) that is linearly related to the fixed and random effects as

$$\begin{aligned}\eta &= \gamma_0 + \gamma_{V1}\text{V1}_s + \gamma_{V2}\text{V2}_s \\ &\quad + \gamma_{MCAV3}\text{MCAV3}_s + \gamma_{MCAV5}\text{MCAV5}_s \\ &\quad + \gamma_{OM}\text{OM}_s + \gamma_{TPI}\text{TPI}_s + \gamma_{CTI}\text{CTI}_s + u_s,\end{aligned} \tag{3}$$

using the logit link. Throughout fitting of candidate sub-models, Eq. 3 (specifying sub-model "model2") is modified iteratively to examine various covariates.

The fixed effects include the first and second decomposed soil variables from the PCA ($\text{V1}_s$) and ($\text{V2}_s$), the third and fifth decomposed soil variables from the MCA ($\text{MCAV3}_s$) and ($\text{MCAV5}_s$), Organic Mass ($\text{OM}_s$), the Topographic Position Index ($\text{TPI}_s$), and the Compound Topographic Index ($\text{CTI}_s$). The random effect ($u_s$) follows a Besag formulation (Besag 1975):

$$u_i | u_j, i \neq j, \tau \sim N\left(\frac{1}{m_i}\sum_{i\sim j} u_j, \frac{1}{m_i \tau}\right) \tag{4}$$

where $N$ is the normal distribution with mean $1/m_i \cdot \sum_{i\sim j} u_j$ and variance $1/(m_i \cdot \tau)$ where $m_i$ is the number of neighboring cells of cell $i$ and $\tau$ is the precision; $i \sim j$ indicates cells $i$ and $j$ are neighbors.

Vague Gaussian priors with known precision are assigned to the $\gamma$'s. The priors and the likelihood are combined with Bayes rule to obtain the posterior distributions for the model parameters. The integrals cannot be solved analytically; therefore, the method of INLA is used as a fast alternative to MCMC simulation for models that have a latent Gaussian

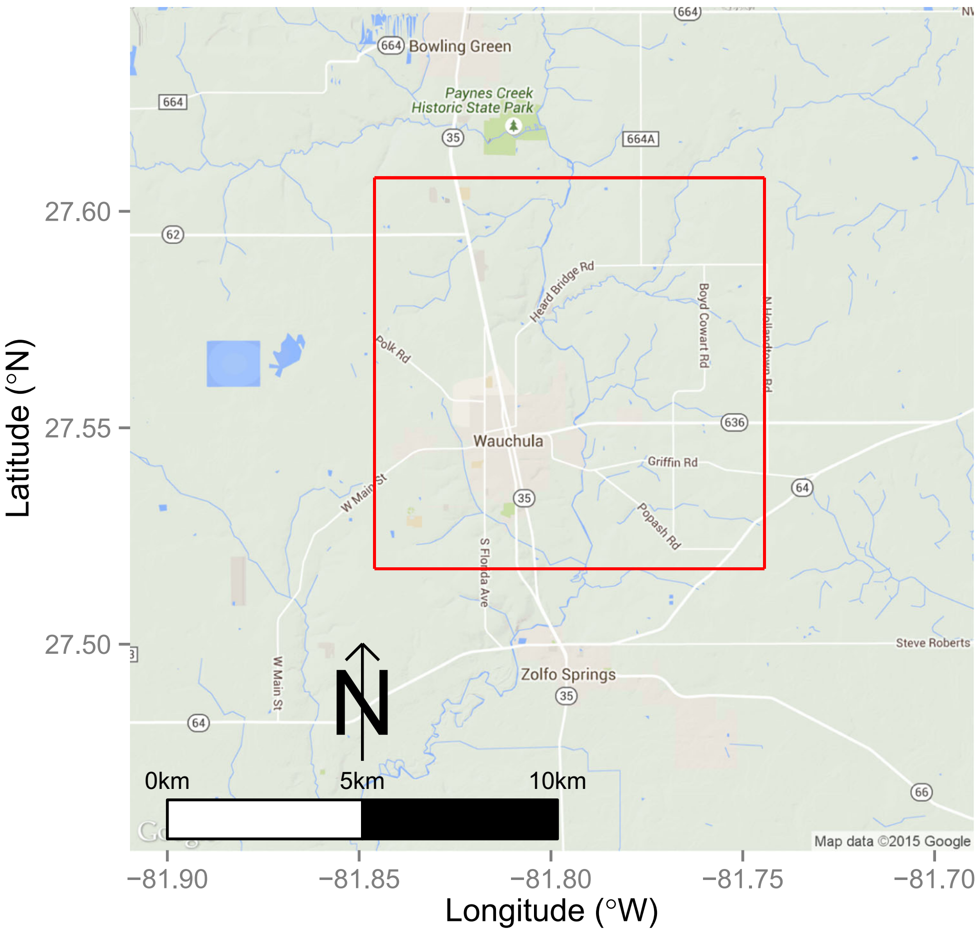

**Fig. 3** Domain 5 located in southwest Florida. The 100 km$^2$ domain is found in Hardee County, Florida

structure (Rue et al. 2009). This is done with functions from the **r-INLA** package (Rue et al. 2014).

The variables developed during the preceding preliminary analysis are evaluated for possible collinearity prior to initiating construction of the spatial model. Collinearity diagnostics for independent variables as described by Belsley et al. (1980) are implemented using the **perturb** package (Hendricks and Pelzer 1991). Initial conditioning of the variable matrix produces an overall condition index score of 28.45 and individual decomposition proportions less than 0.500. These values fall below the maximum thresholds proposed by Belsley et al. (1980), suggesting that collinearity will not bias the model.

To quantify latent structural processes, the model is first fitted (sub-model "model0") with the random-effect term. The random-effect term quantifies spatial dependencies through incorporation of the adjacency graph. The adjacency graph provides a spatial index and neighbors list for all regions within the study domain. Having defined neighbor contiguity during preliminary analysis as being limited to a minimum of two points, the primary domain includes 40,000 regions, with an average of 7.94 links per region. The four corners of the square domain are found to be the least connected regions with three links each.

Next, the fixed covariates identified during preliminary analysis as being significant are added (Sub-model "model1"). Sub-model "model2" is then constructed by retaining the non-zero covariates resulting from model1. To investigate the influence of the random-effect term, sub-model model2 is re-fit without the random-effect term (sub-model "model3").

Posterior distributions for the candidate models are re-approximated to ascertain the Watanabe-Akaike information criteria (WAIC) and log-conditional predictive ordinance (LCPO), which are methods of cross-validated skill. These statistics measure the relative quality of the model given the available data and both are scaled such that the lower the value, the better the model. After fitting all models, comparison of the WAIC and LCPO and consideration of non-significant covariates reveal that sub-model model2

**Table 1** Comparison of model results for the primary domain

| SUB-MODEL | WAIC | LCPO | COVARIATES |
|---|---|---|---|
| model0 | 6936.79 | 0.097 | u (Spatial structure only) |
| model1 | 5664.46 | 0.077 | V1 + V2 + MCAV3 + MCAV5 + AWC + OM + SAVI + CTI + TPI + u |
| model2 | 5667.77 | 0.077 | V1 + V2 + MCAV3 + MCAV5 + OM + CTI + TPI + u |
| model3 | 15329.81 | 0.192 | model2 covariates excluding spatial structure |

Watanabe-Akaike information criteria (WAIC) and log of the conditional predictive ordinance (LCPO)

is the best performing model. Although model1 produces a slightly less WAIC than does model2 (5664.46 and 5667.77 respectively), two of model1's covariates (AWC and SAVI) are found to be non-significant as determined by the credible interval and the LCPO is found to be identical. Four candidate models are fitted in total for the primary domain, these are summarized with WAIC and LCPO in Table 1.

To undertake formal prediction of wetland presence for the primary domain, several steps are required. Firstly, all data used in fitting the best performing sub-model (model2) is duplicated. Next, the response variable of the duplicate dataset is set to "NA" and then re-combined with the original data as required by the **r-INLA** package. The combined dataset, now twice the length used for initial fitting, is re-fit using the model specification from model2. This process leverages model fixed−effects to perform prediction of wetland presence while controlling for spatial processes.

As a means of gauging predictive accuracy, results for spatial and non−spatial models are compared using a Brier Score. The Brier score is a proper score function that measures the accuracy of probabilistic predictions for binary outcomes and is comparable to the mean squared error. The Brier Score is calculated as the sum of the differences of predicted probabilities of wetland presence and the value one ($p$) for all cells identified as having wetlands by the land cover training data used during pre-processing. Overall model accuracy can in-turn be estimated by a scaled Brier Score given as $Brier_{scaled} = (1 - Brier/Brier_{max}) \times 100\ \%$ , where $Brier_{max} = \bar{p} \times (1 - \bar{p})$. The scaled Brier Score measures model accuracy over the more intuitive range from 0 % to 100 % such that the lower the score, the more accurate the prediction. Applied in this manner, the Brier Score is comparable to Pearson's $R^2$ statistic (Hu et al. 2006).

**Table 2** Dimension descriptions for the V1 and V2 decomposed soil variables retained for model fitting

| Soil Attribute | r (V1) | r (V2) | Attribute Description |
|---|---|---|---|
| aws0100wta | −0.7943 | 0.4989 | Available water storage to 100 cm |
| aws0150wta | −0.6023 | 0.6184 | Available water storage to 150 cm |
| aws025wta | −0.2834 | 0.7984 | Available water storage to 25 cm |
| aws050wta | −0.5728 | 0.7017 | Available water storage to 50 cm |
| elev_h | 0.6157 | 0.3738 | Elevation (highest point) |
| elev_l | 0.4961 | 0.5754 | Elevation (lowest point) |
| elev_r | 0.6341 | 0.4167 | Elevation (center point) |
| niccdcdpct | 0.6306 | −0.2375 | Non-irrigated capability class (%) |
| slope_h | 0.9264 | 0.2967 | Slope (highest point) |
| slope_r | 0.6306 | −0.2375 | Slope (center point) |
| slopegradd | 0.8584 | 0.3853 | Gradient of dominant component |
| slopegradw | 0.8584 | 0.3853 | Weighted average slope all components |
| wtdepannmi | 0.7112 | 0.1018 | Shallowest depth wet soil (Anytime) |
| wtdepaprju | 0.6035 | −0.1523 | Shallowest depth wet soil (April - June) |

The Soil Attribute column provides the soil attribute designation as assigned by the NRCS, the center columns provide the Pearson correlation coefficient (r) for V1 and V2, and the final column provides a description of the soil variable. All listed soil attributes were found to be significant ($p$-value $< 0.001$) in describing the dimension

**Table 3** Dimension descriptions for the MCAV3 and MCAV5 decomposed soil variables retained for model fitting

| Soil Attribute | $R^2$ (MCAV3) | $R^2$ (MCAV5) | Attribute Description |
|---|---|---|---|
| muname | 1.0000 | 1.0000 | Mapping Unit Name |
| foragesuit | 0.9897 | 0.9300 | Forage Suitability Group |
| compname | 0.9336 | – | Component name |
| wlwetland | 0.9003 | 0.6711 | Wetland wildlife suitability |
| drclasswet | 0.8946 | 0.7141 | Drainage (wettest) |
| drclassdcd | 0.8921 | 0.7462 | Drainage (dominant) |
| wlwetplant | 0.8653 | 0.7310 | Suitability for wetland plants |
| wlshalloww | 0.8553 | – | Wetland wildlife suitability (shallow) |
| engdwbll | 0.8314 | – | Basement rating (all components) |
| englrsdcd | 0.8230 | – | Rating for roads or streets |
| hydgrpdcd | 0.7779 | – | Hydrologic group runoff potential |
| engdwobdcd | 0.7672 | – | Basement rating (dominant component) |
| brockdepmi | 0.7425 | 0.7230 | Depth to paralithic or lithic layer |
| nirrcapscl | 0.6839 | – | Non-irrigated soils class |
| niccdcd | 0.6615 | – | Non-irrigated soils (average) |
| nirrcapcl | 0.6615 | – | Non-irrigated soils (broadest class) |
| localphase | 0.6564 | 0.7840 | Component phase criterion |
| pondfreqpr | 0.6156 | – | Subject to inundation at surface (%) |
| hydclprs | 0.5992 | – | Hydric soil classification |
| erocl | 0.5759 | – | Class of accelerated erosion |
| hydricon | 0.5632 | – | Natural condition of the soil |
| runoff | – | 0.5804 | Runoff potential class |
| corsteel | 0.5263 | – | Susceptibility of steel to corrosion |

The Soil Attribute column provides the soil variable designation as assigned by the NRCS, the center columns provide the coefficient of determination ($R^2$) for MCAV3 and MCAV5, and the final column provides a description of the soil variable. All listed soil attributes were found to be significant ($p$-value $<$ 0.001) in describing the dimension

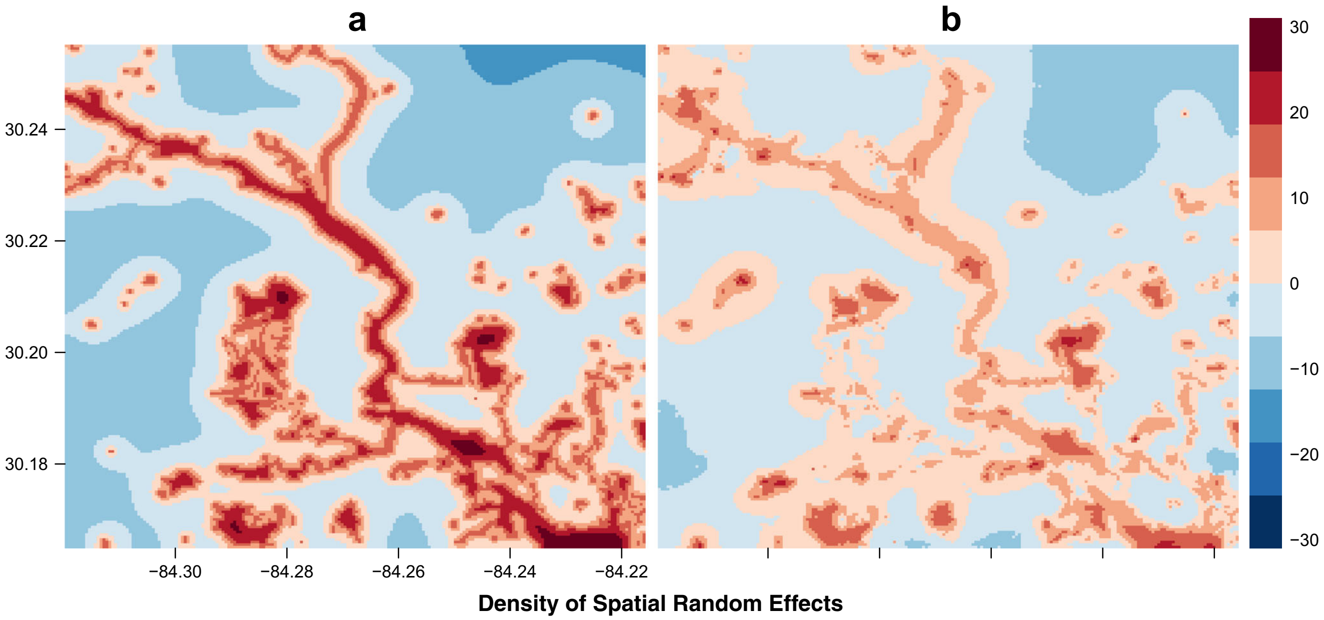

**Fig. 4** Spatially structured random effects for Domain 1. **a** Structured random effects present prior to adding fixed effects (model0). **b** Structured random effects after accounting for fixed effects (*model3*). Percentages indicate change above (*warm colors*) and below (*cool colors*) the domain mean density

## Results

Results from the preliminary analysis indicate that four decomposed soil variables are significant predictors of wetland presence as determined by the credible interval. These variables are retained for model fitting and include the V1 and V2 variables resulting from the PCA and the MCAV3 and MCAV5 variables from the MCA. The continuous soil attributes listed in Table 2 were found to be significantly correlated ($p$-value $< 0.001$) to both V1 and V2 and are displayed with Pearson correlation coefficient (r). Table 3 provides the categorical soil attributes that best describe MCAV3 and MCAV5 with corresponding coefficient of determination ($R^2$). For the categorical attributes, an ANOVA with one factor is conducted for each dimension by regressing the categorical attributes onto the coordinates of the individuals and a F-test is performed to evaluate variable influence on the attribute (Lê et al. 2008).

Initial fitting of the model with the spatial random-effect term in the absence of fixed effects results in a map of smoothed densities relative to the domain mean [Fig. 4a]. Warm colors indicate cells with a density that exceeds the mean density for the domain and cool colors highlight cells where densities fall below the mean. On the whole, the map denotes several areas of elevated density. A linear pattern extends from near the northwest corner and traces a path diagonally across the domain to the southeast corner concomitant with the Wakulla River floodplain. The most concentrated area of density is located in the southeast corner of the domain. Geographically, the southeast corner approaches the confluence of the Wakulla River with the Gulf of Mexico. The southwest and northeast corners demarcate comparatively less-dense areas.

Adding the fixed effects identified with the selected model (model2) allow for remapping of the random-effect while controlling for the soil and hydrologic fixed effects [Fig. 4b]. Regions of elevated density still evident in Fig. 4b provide an estimate for the structural processes that remain after accounting for the indicators of wetland presence specified by the selected model. In comparison to the raw random-effects shown in Fig. 4a, the corrected map for sub-model model2 depicted in Fig. 4b displays a distinct decrease in density across the domain as a whole and particularly along the linear pattern associated with the Wakulla River.

**Table 4** Summary of posterior mean and 95 % Credible Interval for sub-model model2 fixed effects

| | Mean | Quant0.025 | Quant0.975 |
|---|---|---|---|
| (Intercept) | 12.5214 | −13.9087 | −11.4532 |
| V1 | 0.4453 | 0.0533 | 0.8399 |
| V2 | −1.9288 | −2.2383 | −1.6435 |
| MCAV3 | 4.4632 | 4.0206 | 4.9758 |
| MCAV5 | 0.2411 | 0.0313 | 0.4519 |
| OM | −0.2197 | −0.3630 | −0.0765 |
| CTI | 1.8334 | 1.6285 | 2.0645 |
| TPI | −0.4455 | −0.5909 | −0.3013 |

Posterior densities and the corresponding credible intervals for model fixed effects are summarized in Table 4. The first decomposed soil variable (V1) has a posterior mean of 0.4453[(0.0533, 0.8399) 95 % credible interval]. This translates to a 60.95 % $[(1/(1 + \exp(-0.4453))) \times 100\,\%]$ increase in the probability of wetland presence for each whole number increase in the V1 value while holding all other fixed effects constant. Comparatively, the posterior mean of the second soil variable (V2) indicates a 12.69 % increase in the probability of wetland presence for each whole number decrease in V2. The decomposed soil variables resulting from the MCA analysis of categorical data (MCAV3 and MCAV5) indicate that the probability of wetland presence increases 98.86 % and 56.00 % respectively for each unit increase while holding other effects constant. The posterior mean of OM indicates that for each unit decrease in organic mass, the probability of wetland presence increases 44.53 % holding others constant. For each whole unit decrease in the Topographic Position Index (TPI), the probability of wetland presence increases 39.04 % with other effects constant. The posterior mean of the Compound Topographic Index (CTI) translates to a 86.22 % increase in wetland presence for each increase in the CTI index value while holding other fixed effects constant.

To gauge predictive accuracy, results for spatial and non−spatial models are compared using a scaled Brier Score. The lower the Brier score, the better calibrated the model. Resulting scaled Brier Scores indicate that the spatially explicit model produces substantially increased predictive accuracy over its non−spatial equivalent across all domains. In the case of the primary domain (Domain 1), the spatial model achieves a scaled Brier Score of 90.7 % compared to a 75.8 % for the non−spatial model. Results for the three predicted domains are summarized in Table 5.

**Table 5** Summary of scaled Brier Scores

| Domain | Spatial | Non-Spatial |
|---|---|---|
| Domain 1 | 90.7 % | 75.8 % |
| Domain 4 | 86.6 % | 82.1 % |
| Domain 5 | 86.3 % | 76.9 % |

After constructing a model for the primary domain, model robustness is evaluated by changing the spatial resolution, areal extent, and geographic location of the study area. To examine the influence of resolution, the primary domain is reevaluated using cells with a 30 m edge length, and then with cells having a 70 m edge length. The 30 m resolution is selected as it is the resolution of the SAVI covariate. Figure 5a depicts the posterior distributions of model fixed effects at 30 m and 70 m resolutions relative to the 50 m resolution used when fitting the original model. Generally, refining spatial resolution results in an increase to the absolute value of the TPI, MCAV3, and OM fixed effects while a decrease in absolute value is noted for CTI. Although the responses of individual covariates differ, all of model2's fixed effects remain significant at both the 30 m and 70 m resolutions.

Doubling the domain areal extent from 100 km$^2$ to 200 km$^2$ does not result in the loss of any significant fixed effects [Fig. 5b]. That is, all covariates remain significant at both the 200 km$^2$ and 300 km$^2$ areal extents. No general pattern is noted in regard to the influence of extent on the absolute value of fixed effects as each of the covariates uniquely respond to the change; however, a sign change is apparent for several covariates (V1, MCAV3, and MCAV5).

To evaluate model performance in other physical landscapes, the selected model from the primary domain (Domain 1) is applied for two other locations in Florida. Firstly, model fitting and prediction of wetland presence is carried out for a different 100 km$^2$ area (Domain 4) in the same physiographic province as the primary domain and secondly for a 100 km$^2$ area (Domain 5) in peninsular Florida. Echoing the prediction procedure for the primary

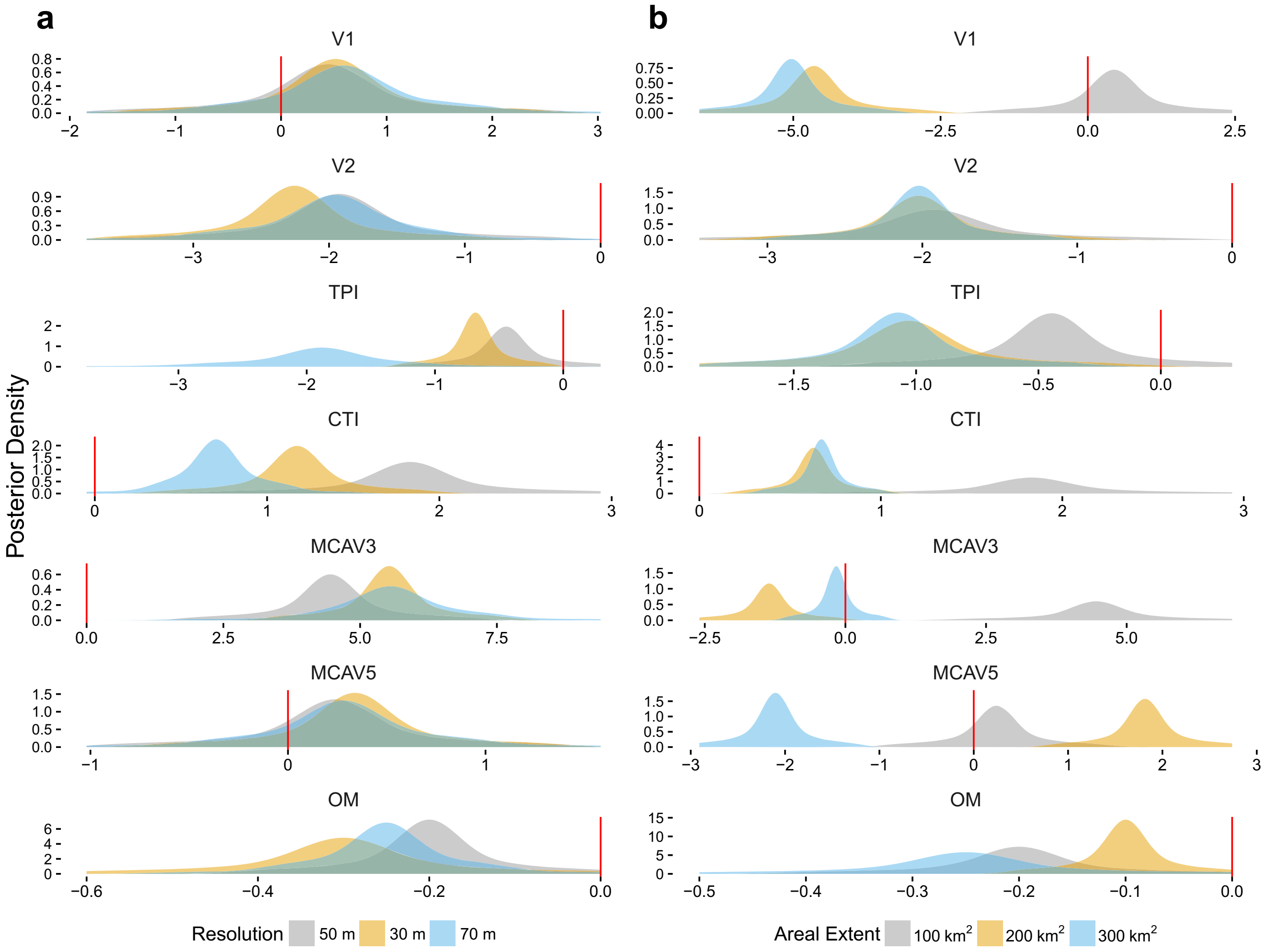

**Fig. 5** Comparison of posterior distributions. **a** Posterior distributions of fixed effects at 30 m, 50 m, and 70 m spatial resolutions. **b** Posterior distributions of fixed effects at extents of 100 km$^2$, 200 km$^2$, and 300 km$^2$

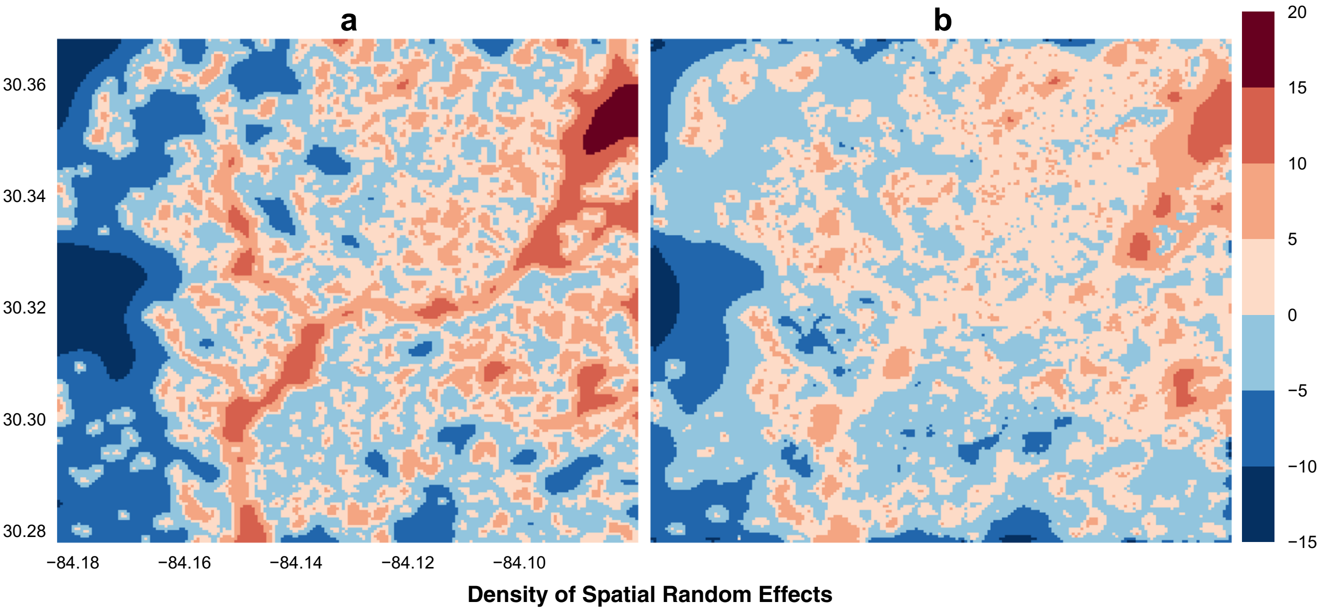

**Fig. 6** Spatially structured random effects for Domain 4. **a** Structured random effects present prior to adding fixed effects (*model0*). **b** Structured random effects after accounting for fixed effects (*model2*). Percentages indicate change above (*warm colors*) and below (*cool colors*) the domain mean density

domain, both spatial and non−spatial models are implemented to better understand the contribution of the spatial random-effect term.

As with the primary study area, fitting of the fourth and fifth domains with the spatial random-effect term in the absence of fixed effects reveals regions of elevated density.

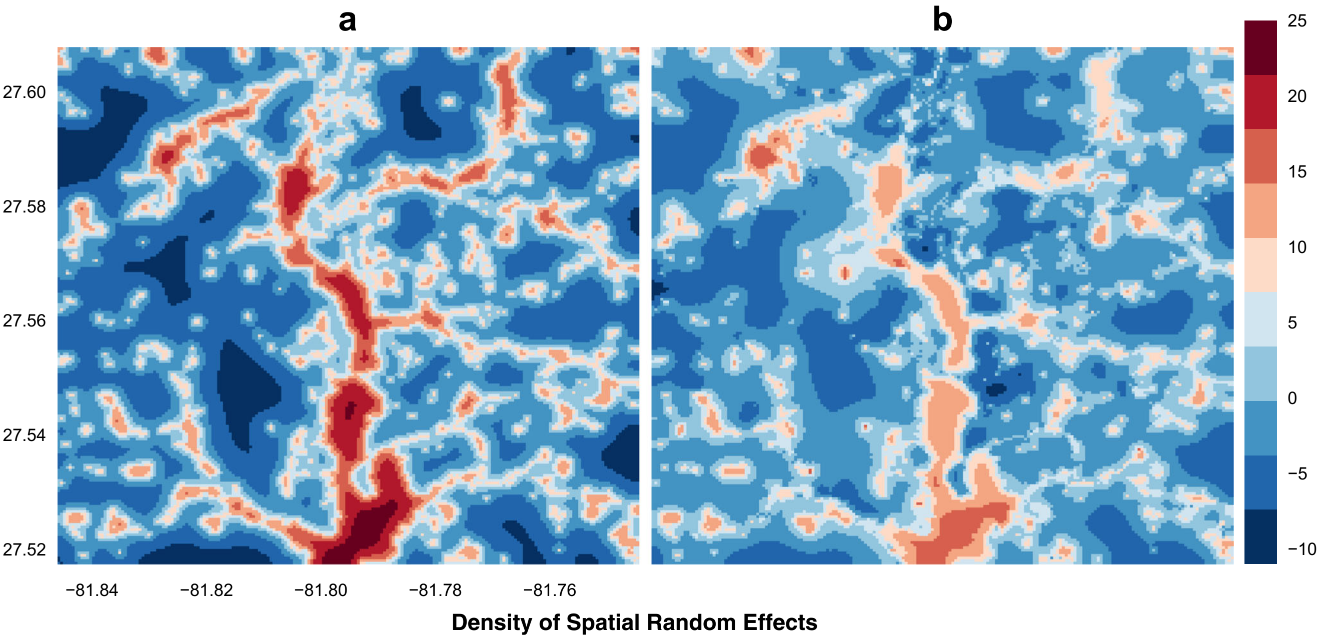

**Fig. 7** Spatially structured random effects for Domain 5. **a** Structured random effects present prior to adding fixed effects (*model0*). **b** Structured random effects after accounting for fixed effects (*model3*). Percentages indicate change above (*warm colors*) and below (*cool colors*) the domain mean density

In both cases, subsequent control of fixed effects reduces total domain density. Comparison maps of random-effects for the fourth domain are provided in Fig. 6 and maps for the fifth domain are shown in Fig. 7.

The wetland maps produced for each of the five study domains identify probable wetlands at locations not identified as wetlands by the National Wetlands Inventory (NWI). Review of aerial imagery for locations with a relatively high predicted probability of presence (e.g. probability greater than 0.75) verify evidence of wetlands in the majority of these cases. Figure 8 compares model predicted wetlands for the primary domain (Domain 1) to those identified by the NWI for the same area. The map located at the bottom (C) of Fig. 8 displays the difference between model predicted wetlands and those identified by the NWI. This illustration was created by reclassifying locations identified as wetlands by the NWI to 0 %. Comparison maps for Domain 4 and Domain 5 are provided in Figs. 9 and 10 respectively. To aid in interpretation of Fig. 10, a Flood Insurance Rate Map (FIRM) produced by the Federal Emergency Management Agency (FEMA) is also provided [Fig. 10c] to illustrate the base floodplain (i.e., 100 year floodplain, Flood Hazard Zones "A" and "AE"). FIRM data is freely available from the Federal Emergency Management Agency [https://msc.fema.gov/portal]. Assuming locations with a model predicted probability of wetland presence greater than the arbitrary threshold of 0.50 to represent realized wetlands, models for all domains identify a greater area of wetland (number of cells) than does the NWI. Beyond comparison of total area, model predicted wetlands across all domains include numerous flow ways, transitional areas, and ecotone gradients outside of areas identified by the NWI as wetlands. Typically, predicted probabilities for transitional areas were between 0.30 and 0.49. Because the probability of wetlands at these locations fall below the arbitrary threshold of 0.50, these extents are not included in the areal comparison provided above for illustrative purposes.

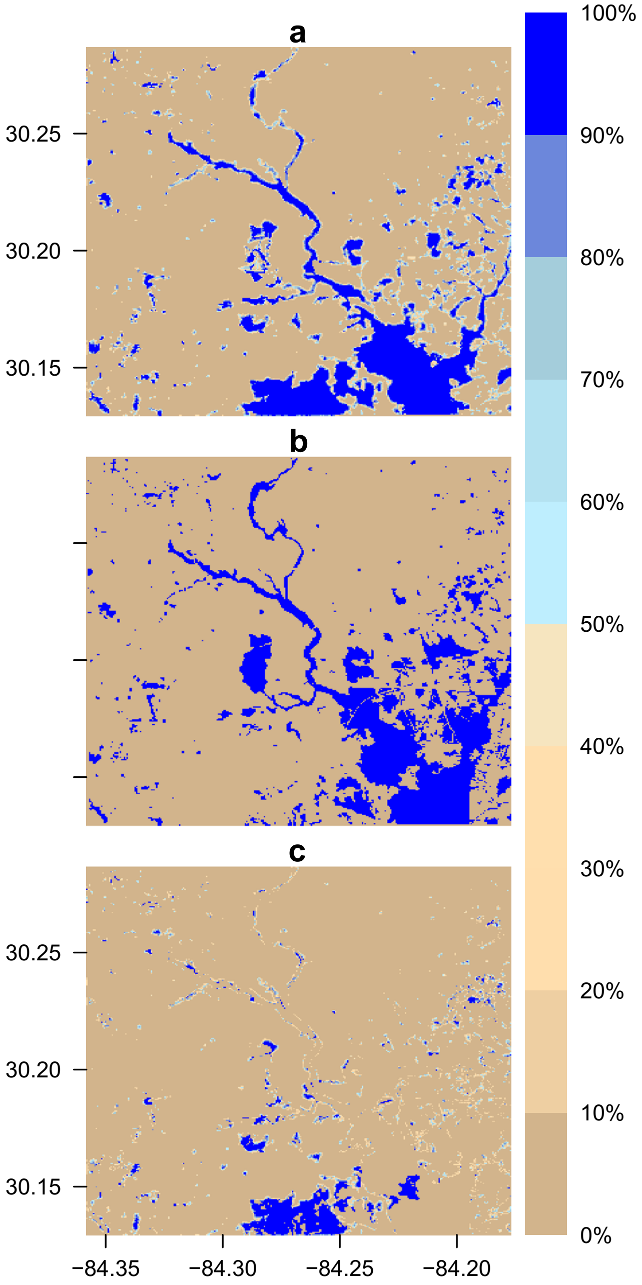

**Fig. 8** Comparison of model predicted wetlands for Domain 1 (*northwest Florida*) with wetlands identified by the NWI. Top map (**a**) depicts the extent of wetlands as predicted by the model. The map at center (**b**) displays wetland extents as identified by the NWI. The map at bottom (**c**) displays wetland extents identified by the model but not identified by the NWI (e.g. NWI wetlands reclassified as 0 %). All maps represent approximately 100 km$^2$ and are composed of square cells with an edge length of 50 m

## Discussion

The identification and inventory of wetlands is essential to natural resource management. To be effective in these endeavors, it is critical that the procedures used to detect and document wetlands be time efficient, accurate, and economical. Aerial photographic interpretation of land cover by individual human analysts necessitates hours of assessment, introduces human error, and fails to include the best available soils and hydrologic data. Furthermore, photographic interpretation results in products that assume the patch matrix perspective. Under the patch matrix view, real-world landscape gradients are reduced to dichotomous patches of wetland presence or absence. The current study leverages probabilistic models to predict wetland presence as a continuous gradient (from not likely to certain) with explicit consideration of spatial processes.

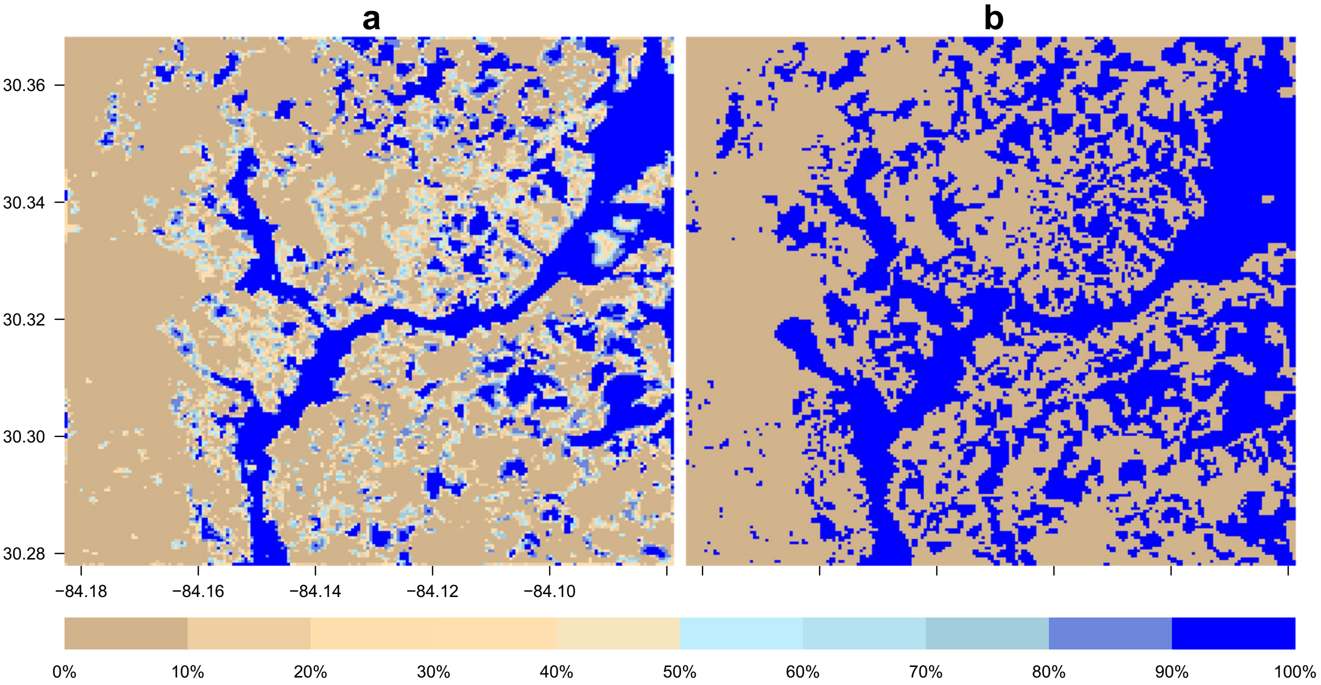

**Fig. 9** Comparison of model predicted wetlands for Domain 4 (*northwest Florida*) with wetlands identified by the NWI. Map (**a**) depicts the extent of wetlands as predicted by the model. Map (**b**) displays wetland extents as identified by the NWI. All maps represent approximately 100 km$^2$ and are composed of square cells with an edge length of 50 m

The presented model incorporates freely available elevation, land use, and soils data and is executed as a single R-code script. In the experience of the authors, identification of wetlands over a 100 km$^2$ area via photographic interpretation necessitates approximately 40 work hours by an experienced human analyst. By comparison, the presented model can evaluate 100 km$^2$ at 50 meter resolution in approximately 50 minutes (Intel Core i7 4712HQ x4 2.3GHz 16GB) while incorporating ancillary data and accounting for latent spatial processes. Model results demonstrate an ability to consistently capture training data derived from heads-up assessment with greater than 90 % accuracy as judged by the scaled Brier Score, to out perform non-spatial linear models, and to identify wetland extents, ecotones, and hydrologic connections not included in either the training data or the NWI.

Although, the provided model is reasonably robust to changes in resolution, areal extents between 100 km$^2$ and 300 km$^2$, and region-specific physical conditions, it is advisable that model fitting procedures be repeated for each unique study domain as a means of calibrating the model to local biophysical conditions. It may also be the case that the environmental covariates explored in the current study are not available for some locations outside of the United States, or that other topographic or vegetative indices are found to offer greater explanatory power due to region-specific characteristics. That is, caution should be used when exporting the best performing model from one region to another. The application of the best performing model from the primary domain to other locations in this study was for the sole purpose of evaluating model sensitivity. Others are encouraged to utilize the presented modeling approach, to apply a gradient-based perspective, to leverage spatial structure towards prediction, and to widely explore other potential environmental covariates.

Finally, several aspects of this study warrant further investigation and it is the authors' hope that the presented work inspires exploration by other researchers. For example, although decomposition of soils data facilitated the development of several strong predictors of wetland presence, results from the sensitivity analysis bring to light issues in regard to consistent use of such an approach. Among these are issues relating to the shift of a given dimension between positive and negative signs at differing spatial extents, changes in the magnitude of effect sizes over different study domains, and the overall repeatably of what is essentially a data mining technique. Many of these potentially confounding factors were not fully explored in this study, as the presented model framework is motivated by prediction rather than explanation or causality.

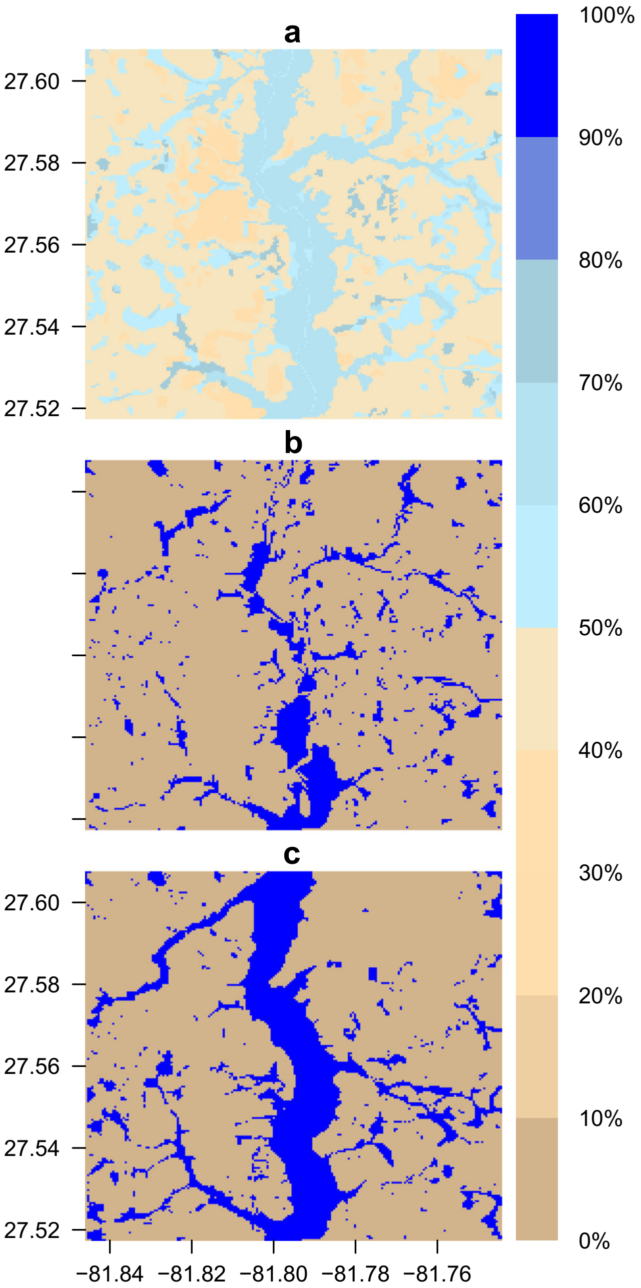

**Fig. 10** Comparison of model predicted wetlands for Domain 5 (*southwest Florida*) with wetlands identified by the NWI and FEMA FIRM High Risk flood hazard areas (A and AE). Top map (**a**) depicts the extent of wetlands as predicted by the model. Map at center (**b**) displays wetland extents as identified by the NWI. Map at bottom (**C**) displays FEMA FIRM "High Risk" flood Zones A and AE (i.e., base floodplain). All maps represent approximately 100 km$^2$ and are composed of square cells with an edge length of 50 m

**Acknowledgments** This study was funded in part by an U.S. Environmental Protection Agency Wetland Program Development Grant (00D14313) and support from the Florida Department of Environmental Protection.

## References

Adam E, Mutanga O, Rugege D (2010) Multispectral and hyperspectral remote sensing for identification and mapping of wetland vegetation: A review. Wetlands Ecology and Management 18(3):281–296. doi:10.1007/s11273-009-9169-z

Barbier EB (2011) Wetlands as natural assets. Hydrological Sciences Journal 56(8):1360–1373. doi:10.1080/02626667.2011.629787

Beaudette D, Skovlin J, Roecker S (2016) soilDB: Soil Database Interface. https://CRAN.R-project.org/package=soilDB, r package version 1.8

Belsley DA, Kuh E, Welsch RE (1980) Identifying influential data and sources of collinearity. Wiley, New Jersey

Besag J (1975) Statistical analysis of non-lattice data. Journal of the Royal Statistical Society Series D (The Statistician) 24:179–195

Bivand R, Piras G (2015) Comparing Implementations of Estimation Methods for Spatial Econometrics. Journal of Statistical Software 63(18)

Christensen N (1996) The Report of the Ecological Society of America Committee on the Scientific Basis for Ecosystem Management. Ecological Applications 6(3):665–691. doi:10.2307/2269460

Dahl TE (2011) Status and Trends of Wetlands in the Conterminous United States 2004 to 2009. US Department of the Interior; Fish and Wildlife Service, Washington, DC (November): 108

Dvorett D, Bidwell J, Davis C, DuBois C (2012) Developing a hydrogeomorphic wetland inventory: Reclassifying national wetlands inventory polygons in geographic information systems. Wetlands 32(1):83–93. doi:10.1007/s13157-011-0247-7

Elsner JB, Fricker TW, HMC, Humphreys J, Jung J, Gredzens C (2016) The relationship between elevation roughness and tornado activity: A spatial statistical model fit to data from the Central Great Plains. Journal of Applied Meteorology and Climatology 55(4):849–859. doi:http://dx.doi.org/10.1175/JAMC-D-15-0225.1

Evans JS, Cushman SA (2009) Gradient modeling of conifer species using random forests. Landscape Ecology 24(5):673–683. doi:10.1007/s10980-009-9341-0

Finlayson CM, Davidson NC, Spiers AG, Stevenson NJ (1999) Global wetland inventory – current status and future priorities. Marine and Freshwater Research 50(8):717. doi:10.1071/MF99098

Fish and Wildlife Service (2009) Status Report for the National Wetlands Inventory Program: 2009. Tech. rep., U.S. Fish and Wildlife Service, Arlington, Virginia 22203

Galbraith JM, Donovan PF, Smith KM, Zipper CE (2003) Using Public Domain Data To Aid in Field Identification of Hydric Soils. Soil Science 168(8):563–575. doi:10.1097/01.ss.0000085049.25696.84

Hendricks J, Pelzer B (1991) Collinearity involving ordered and unordered categorical variables. RC33 conference in Amsterdam:1–17

Hoeting JA (2009) HoetingThe importance of accounting for spatial and temporal correlation in analyses of ecological data. Ecological applications 19(3):574–577

Hu B, Palta M, Shao J (2006) Properties of R(2) statistics for logistic regression. Statistics in Medicine 25(8):1383–95. doi:10.1002/sim.2300

Kahle D, Wickham H (2013) ggmap: Spatial Visualization with ggplot2. The R Journal 5(1):144–161

Lang M, McDonough O, McCarty G, Oesterling R, Wilen B (2012) Enhanced detection of wetland-stream connectivity using lidar. Wetlands 32(3):461–473. doi:10.1007/s13157-012-0279-7

Lang M, McCarty G, Oesterling R, Yeo IY (2013) Topographic metrics for improved mapping of forested wetlands. Wetlands 33(1):141–155. doi:10.1007/s13157-012-0359-8

Lausch A, Blaschke T, Haase D, Herzog F, Syrbe R, Tischendorf L, Walz U (2015) Understanding and quantifying landscape structure

– A review on relevant process characteristics, data models and landscape metrics. Ecological Modelling 295(August):31–41. doi:10.1016/j.ecolmodel.2014.08.018

Lê S, Josse J, Husson F (2008) FactoMineR : An R package for multivariate analysis. J Stat Softw 25(1):1–18. doi:10.1016/j.envint.2008.06.007

Martin GI, Kirkman LK, Jeffrey HC (2012) Mapping geographically isolated wetlands in the dougherty plain, georgia, USA. Wetlands 32:149–160. doi:10.1007/s13157-011-0263-7

Masek J, Vermote E, Saleous N, Wolfe R, Hall F, Huemmrich F, Gao F, Kutler J, Lim TK (2006) A Landsat surface reflectance data set for North America. IEEE Geoscience and Remote Sensing Letters 3:68–72

Mccauley LA, Jenkins DG (2005) GIS-Based Estimates of Former and Current Depressional Wetlands in an Agricultural Landscape Published by : Ecological Society of America GIS-BASED ESTIMATES OF FORMER AND CURRENT DEPRESSIONAL WETLANDS IN AN AGRICULTURAL LANDSCAPE. Ecological Applications 15(4):1199–1208

McGarigal K (2005) The Gradient Concept of Landscape Structure. In: Issues and perspectives in landscape ecology. Cambridge University Press, pp 112–119

McGarigal K, Tagil S, Cushman SA (2009) Surface metrics: An alternative to patch metrics for the quantification of landscape structure. Landscape Ecology 24(3):433–450. doi:10.1007/s10980-009-9327-y

Meixler MS, Bain MB (2010) Landscape scale assessment of stream channel and riparian habitat restoration needs. Landscape and Ecological Engineering 6(2):235–245. doi:10.1007/s11355-010-0103-6

Mitsch WJ, Gossilink JG (2000) The value of wetlands: Importance of scale and landscape setting. Ecological Economics 35(1):25–33. doi:10.1016/S0921-8009(00)00165-8

Moore ID, Gessler PE, NG A, Petersen GA (1993) Terrain attributes: estimation methods and scale effects. In: Jakeman JA, Beck MB, Wiley MM (eds) Modeling Change in Environmental Systems, London, pp 189–214

Murphy C, Ogil S, Arp PA (2007) Mapping wetlands: A comparison of two different approaches for New Brunswick, Canada. Wetlands 27(4):846–854. doi:10.1672/0277-5212(2007)27

Pennock D, Bedard-Haughn A, Kiss J, van der Kamp G (2014) Application of hydropedology to predictive mapping of wetland soils in the Canadian Prairie Pothole Region. Geoderma 235-236:199–211. doi:10.1016/j.geoderma.2014.07.008

R Core Team (2016) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rebelo LM, Finlayson CM, Nagabhatla N (2009) Remote sensing and GIS for wetland inventory, mapping and change analysis. Journal of Environmental Management 90(7):2144–2153. doi:10.1016/j.jenvman.2007.06.027

Reif M, Frohn RC, Lane CR, Autrey B (2009) Mapping Isolated Wetlands in a Karst Landscape: GIS and Remote Sensing Methods. GIScience & Remote Sensing 46(2):187–211. doi:10.2747/1548-1603.46.2.187

Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society Series B: Statistical Methodology 71(2):319–392. doi:10.1111/j.1467-9868.2008.00700.x

Rue H, Martino S, Lindgren F, Simpson D, Riebler A, Krainski ET (2014) INLA: Functions which allow to perform a full Bayesian analysis of structured (geo-)additive models using Integrated Nested Laplace Approximation. R package version 00-1383402327

Turner MG (1989) Landscape Ecology: The Effect of Pattern on Process. Annual Review of Ecology and Systematics 20(1):171–197. doi:10.1146/annurev.es.20.110189.001131

Veihmeyer F, Hendrickson AH (1927) The relation of soil moisture to cultivation and plant growth. 1st Intern Congr Soil Sci 3:498–513

Wiens JJA (1989) Spatial scaling in ecology. Functional Ecology 3(4):385–397. doi:10.2307/2389612

Wu J, Hobbs R (2002) Key issues and research priorities in landscape ecology: An idiosyncratic synthesis. Landscape Ecology 17(3):355–365. doi:10.1023/A:1020561630963

Wu J, Hobbs R (2007) Scale and scaling : a cross-disciplinary perspective. In: Key Topics in Landscape Ecology, Cambridge University Press, chap 7, pp 115–142

Wu J, Loucks O (1995) From balance of nature to hierarchical patch dynamics. The Quarterly Review of Biology 70(4):439–466