Weather Simulation Uncertainty Estimation Using Bayesian Hierarchical Models

JIANFENG WANG

Department of Mathematics and Mathematical Statistics, Umeå University, Umeå, Sweden

RICARDO M. FONSECA

Group of Atmospheric Science, Division of Space Technology, Department of Computer Science, Electrical and Space Engineering, Luleå University of Technology, Luleå, Sweden

KENDALL RUTLEDGE

Novia University of Applied Sciences, Vaasa, Finland

JAVIER MARTÍN-TORRES

Group of Atmospheric Science, Division of Space Technology, Department of Computer Science, Electrical and Space Engineering, Luleå University of Technology, Luleå, Sweden, and Instituto Andaluz de Ciencias de la Tierra (CSIC-UGR), Granada, Spain

JUN YU

Department of Mathematics and Mathematical Statistics, Umeå University, Umeå, Sweden

(Manuscript received 18 January 2018, in final form 15 January 2019)

ABSTRACT

Estimates of the uncertainty of model output fields (e.g., 2-m temperature, surface radiation fluxes, or wind speed) are of great value to the weather and climate communities. The traditional approach for the uncertainty estimation is to conduct an ensemble of simulations where the model configuration is perturbed and/or different models are considered. This procedure is very computationally expensive and may not be feasible, in particular for higher-resolution experiments. In this paper, a new method based on Bayesian hierarchical models (BHM) that requires just one model run is proposed. It is applied to the Weather Research and Forecasting (WRF) Model's 2-m temperature in the Botnia–Atlantica region in Scandinavia for a 10-day period in the winter and summer seasons. For both seasons, the estimated uncertainty using the BHM is found to be comparable to that obtained from an ensemble of experiments in which different planetary boundary layer (PBL) schemes are employed. While WRF-BHM is not capable of generating the full set of products obtained from an ensemble of simulations, it can be used to extract commonly used diagnostics including the uncertainty estimation that is the focus of this work. The methodology proposed here is fully general and can easily be extended to any other output variable and numerical model.

1. Introduction

The estimation of the uncertainty of a model prediction is a critical issue that has become increasingly relevant in the last decades with the use of more complex and sophisticated numerical models. When a variable like temperature is measured with a thermometer at a given location and time, the uncertainty that goes with

the measurement can be readily estimated. In the same way, the model predictions at a given grid box and time optimally should come with an uncertainty estimate. The uncertainty of simulated weather and climate variables has been the subject of a significant number of published works. It is inherent to the chaotic nature of the nonlinear equations solved in weather and climate models (Slingo and Palmer 2011) and has important implications for short-term weather forecasts (e.g., Fox and Wikle 2005), future climate change studies

Corresponding author: Jianfeng Wang, jianfeng.wang@umu.se

DOI: 10.1175/JAMC-D-18-0018.1

© 2019 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy](https://www.ametsoc.org/PUBSReuseLicenses) (www.ametsoc.org/PUBSReuseLicenses).

(e.g., Salazar et al. 2016; Overland et al. 2016), ocean (e.g., Bitner-Gregersen et al. 2014), and chemistry (e.g., Malmberg et al. 2008) research and intermodel comparison (e.g., Wang et al. 2016). As discussed in Leutbecher and Palmer (2008), there are two main sources of errors in numerical weather prediction models that result in uncertainties: initial condition errors and model errors. The former arises because the data used to initialize the models are not perfect, with some of these errors amplifying during the simulation and resulting in significant forecast errors. The latter is due to deficiencies in the model's representation of the dynamics and physics of the atmosphere by numerical algorithms. The traditional way to estimate the uncertainty of a model variable is to perform an ensemble of experiments in which one (or more) aspect(s) of the model configuration is varied: for example, different physics and/or dynamics options are selected (e.g., Evans et al. 2012) and the initial and/or boundary conditions are perturbed (e.g., Torn and Hakim 2009). Another commonly followed approach is a multimodel ensemble (e.g., Salazar et al. 2016). The uncertainty sampled with an ensemble of simulations will, of course, depend on how the ensemble is set up. For example, in a multi-physics ensemble, the uncertainty estimate will reflect the sensitivity of the model's response to the different treatments of a given process (e.g., boundary layer dynamics, cumulus convection, or cloud microphysics processes). Another possibility is to perturb the initial and/or boundary conditions with some measurement of uncertainty associated with them. In this case, it will be an indication of how the uncertainty associated with the forcing data propagates to the model response. Philosophically, such ensemble methods are fundamentally different from measurement error and natural variability of variables in nature. Ensemble uncertainty estimation methods quantify the differences in various researchers' implementation and interpretation of nature. We point out that such model ensemble error summaries are conceptually different than natural variability plus measurement uncertainty. These methods are not only time consuming but very computationally demanding; therefore, slowing down the progression into higher spatial resolutions is needed for a more realistic simulation of the atmospheric conditions (e.g., Prein et al. 2015). Providing the uncertainty for a time-dependent spatial field like the 2-m temperature is crucial for a majority of weather and climate applications, not least for climate change projections, as it will help guide policy-makers and stakeholders responsible for developing mitigation strategies. There is a need to develop efficient ways of estimating the uncertainty of a model prediction that uses the least possible amount of

computational resources. In this paper, such an approach is proposed that makes use of a single model run and a statistical model, the Bayesian hierarchical model (BHM).

The vast majority of the statistical models used in numerical weather research are based on multiple linear regression, such as model output statistics (MOS) techniques (Vislocky and Fritsch 1995, 1997) and the Statistical Hurricane Intensity Prediction Scheme (SHIPS; DeMaria and Kaplan 1994; DeMaria et al. 2005). Statistical models based on artificial neural networks (e.g., Kuligowski and Barros 1998; Sideratos and Hatzigeorgiou 2007), linear stochastic processes (e.g., Toth et al. 2000), and Bayesian model averaging (e.g., Raftery et al. 2005) have also been used. The BHM proposed here is a hierarchical model inferred using the Bayesian approach (i.e., the different layers in the BHM are linked through Bayes's rule), which allows both observed data and model parameters to be random variables, resulting in a more realistic and accurate estimation of the uncertainty compared to non-Bayesian approaches (e.g., Haining et al. 2007). In addition, flexible prior information can be incorporated into the Bayesian model: as pointed out by Dormann (2007), prior information (i.e., information that accounts for subjective beliefs or expectations regarding the behavior of a given parameter before it is estimated from the data itself) can be very helpful in modeling spatial autocorrelative effects compared to ordinary nonspatial linear effects. These features of the Bayesian model are behind our choice of this approach for this work.

An example of a problem that has been successfully addressed using a BHM approach is the estimation of the rate of suicide mortality in 32 London boroughs (Blangiardo et al. 2013). The number of suicides in each borough, as well as their location, is known. Assuming that the number of suicides is conditionally independent given its suicide rate, in the sense that the occurrence of one does not affect the probability of a second to occur, the number of suicides y_{s_i} will follow a Poisson distribution:

$$y_{s_i} | \lambda_{s_i} \sim \text{Poisson}(\lambda_{s_i}), \quad (1)$$

where s_i is the location index with $i = \{1, 2, \dots, 32\}$, and λ_{s_i} is the suicide rate at s_i . The parameter $\boldsymbol{\lambda} = \{\lambda_{s_1}, \lambda_{s_2}, \dots, \lambda_{s_{32}}\}$ is modeled by a multivariate Gaussian distribution:

$$\boldsymbol{\lambda} | \boldsymbol{\zeta} \sim \mathcal{N}[\boldsymbol{\zeta}_1, \boldsymbol{\Sigma}(\boldsymbol{\zeta}_2)], \quad (2)$$

which is specified by the parameters $\boldsymbol{\zeta} = \{\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2\}$, where $\boldsymbol{\zeta}_1$ is the mean vector and $\boldsymbol{\Sigma}(\boldsymbol{\zeta}_2)$ is the covariance matrix

indexed by ζ_2 ; ζ is a hyperparameter, as it has to be specified before the method is implemented. An appropriate prior distribution will then be chosen for ζ . Once that is done, and through Bayes's rule, one can derive the posterior distribution,

$$[\lambda, \zeta | \mathbf{y}] \propto [\mathbf{y} | \lambda, \zeta][\lambda | \zeta][\zeta], \quad (3)$$

where $\mathbf{y} = \{y_{s_1}, y_{s_2}, \dots, y_{s_{32}}\}$ is the whole data, and $[\cdot]$ denotes distribution. Finally, $[\lambda, \zeta | \mathbf{y}]$ can be extracted, giving the full distribution of suicide mortality from which statistics such as the mean and standard deviation can be inferred.

One obstacle, however, hinders the use of a Bayesian model: the approximations step. The normalizing constant used in Eq. (3), relevant for model-selection problems but not important for parameter estimation, is typically intractable with the analytical derivation replaced by numerical approximations that are often very costly. The Markov chain Monte Carlo (MCMC; Steinsland and Jensen 2010) method is a key step to make the inference. The basic idea is to draw a sequence of dependent samples from a target density of interest. As the number of draws tends to infinity, the resulting output behaves as if it were a sample from the density of interest. However, as the data become larger and their inherent structure more complex, this procedure becomes very computationally expensive. What is more, for hierarchical models, such as the BHM used here, the convergence of the MCMC is very slow and unpredictable (De Smedt et al. 2015). The estimation of hyperparameters is also challenging for the MCMC. An alternative to the MCMC method for which faster approximations of posterior distributions even for more complex models are possible is the integrated nested Laplace approximation (INLA; Rue et al. 2009) method. The INLA method uses a combination of analytical approximations and numerical integrations to speed up the estimation of the posterior distribution instead of generating a long sequence of dependent samples such as in the MCMC. Methods such as INLA make the Bayesian model increasingly more popular in several applications such as for spatial or spatiotemporal data in imaging analysis and geographical analysis (Cressie and Wikle 2011; Blangiardo et al. 2013). Even with these new approximation methods, the implementation of a Bayesian model still requires large amounts of computational resources [in particular, random-access memory (RAM)] compared to non-Bayesian approaches. In particular, it takes about 50 min and roughly 12 gigabyte (GB) RAM on a computer with an Intel Xeon central processing unit (CPU) E3-1270, version 3, at 3.5 GHz to implement a single

BHM model. Hence, it has to be applied on a high-performance server that can handle big data with a complex dependent structure.

The Weather Research and Forecasting (WRF; Skamarock et al. 2008) Model used in this work is a fully compressible, nonhydrostatic model that uses a terrain-following hydrostatic pressure-based coordinate in the vertical and the Arakawa C grid staggering for horizontal discretization. It has been used in a wide variety of applications including polar meteorology studies (e.g., Hines and Bromwich 2008) and boundary layer research (e.g., Banks et al. 2016). In this work, the uncertainty of the daily mean 2-m temperature is estimated from one WRF run using the BHM. To assess how realistic the WRF-BHM uncertainty estimate is, it will be compared to that obtained from an ensemble of WRF simulations in which different planetary boundary layer (PBL) schemes are employed. The uncertainty that will be targeted in this work is that associated with different treatments of the boundary layer dynamics and representation of the near-surface flow, which play a crucial role in the simulation of the daily mean near-surface air temperature. It is important to stress that WRF-BHM is not capable of generating the full set of products that can be obtained from an ensemble of experiments. An example is the most likely event, such as for precipitation type. Knowing which precipitation type (e.g., rain, snow, ice pellets, or freezing rain) is more likely can be achieved by checking which is predicted by the majority of the ensemble members with precipitation but cannot be obtained from the BHM. As a result, the BHM cannot be considered as a true alternative to ensemble forecasting. Having said that, ensemble products such as the probability of exceedance, box-and-whisker diagrams, background-error covariance for ensemble Kalman filter, and hybrid variational data and uncertainty estimation, the focus of this work, can be obtained in a computationally cheaper and faster way with the BHM. In this paper, the uncertainty estimation is considered for the daily-mean 2-m temperature, but the methodology developed can be easily extended to other fields.

2. Experimental setup

In this study, WRF, version 3.7.1, is forced with the National Centers for Environmental Prediction (NCEP) Climate Forecast System Reanalysis (CFSR; Saha et al. 2010) 6-hourly data (horizontal resolution of $0.5^\circ \times 0.5^\circ$) and is run for one month in the boreal summer (July 2016) and winter (January 2017) seasons. To minimize the accumulation of integration errors, and following Lo et al. (2008), each month's run is broken into three

overlapping 11–12-day periods (e.g., for July 2016, the model is run from 0000 UTC 30 June to 0000 UTC 11 July, from 0000 UTC 10 July to 0000 UTC 21 July, and from 0000 UTC 20 July to 0000 UTC 1 August) with the first day regarded as model spinup.

The physics parameterizations used include the Goddard 6-class microphysics scheme (Tao et al. 1989), the Rapid Radiative Transfer Model for Global Circulation Models (RRTMG) models for both shortwave and longwave radiation (Iacono et al. 2008), the Mellor–Yamada Nakanishi and Niino (MYNN), level 2.5 (Nakanishi and Niino 2006); with Monin–Obukhov surface layer parameterization (Monin and Obukhov 1954); the 4-layer Noah land surface model (Chen and Dudhia 2001); and for cumulus convection the Betts–Miller–Janjić scheme (Janjić 1994) coupled with the precipitating convective cloud scheme described in Koh and Fonseca (2016). A climatological aerosol distribution based on Tegen et al. (1997) is used in the radiation scheme. An interactive prognostic scheme for the sea surface skin temperature (SSKT) based on Zeng and Beljaars (2005) is also added to the model. The lower boundary condition to the SSKT scheme is the 6-hourly SSTs from CFSR, linearly interpolated in time in order to have a continuously varying forcing on the skin layer. The fractional sea ice coverage and sea ice depth are also read from CFSR every 6 h with the sea ice albedo a function of air temperature, skin temperature, and snow (Mills 2011). Gravitational settling of cloud drops in the atmosphere is parameterized as in Duynkerke (1991) and Nakanishi (2000), while cloud water (fog) deposition onto the surface because of turbulent exchange and gravitational settling is treated using the simple fog deposition estimation (FogDES) scheme (Katata et al. 2011).

In addition to the control experiment (model configuration stated above, with the MYNN, level 2.5, PBL scheme), nine additional simulations are performed using different PBL schemes. As the focus of this work is on the 2-m temperature, it makes sense to perturb the PBL scheme as the near-surface fields are highly sensitive to how the boundary layer dynamics are represented. This is particularly true for Scandinavia, which has a very large seasonal contrast: snow-covered terrain, sea ice in the Gulf of Bothnia, and a very short daylight period in winter, contrasting with open water everywhere, snow-free land, and long daylight hours in summer. These sensitivity runs are conducted over a 10-day period at the beginning of January 2017 and July 2016 with a 1-day spinup before. The winter period selected comprises both milder days, when baroclinic systems from the Atlantic prevail, and colder days, when areas of high pressure are in control and rather cold

temperatures, dropping below -30°C in some places, occur. The summer period started with a southerly flow and rather warm temperatures, but the weather conditions became more unstable in the middle and latter parts of the period. Both are representative of the different atmospheric conditions experienced in Scandinavia in the cold and warm seasons. The PBL schemes considered are the Yonsei University (YSU; Hong et al. 2006); asymmetric convective model, version 2 (ACM2; Pleim 2007); quasi-normal scale elimination (QNSE; Sukoriansky et al. 2005); Mellor–Yamada–Janjić (MYJ; Janjić 1994); Bougeault–Lacarrère (BouLac; Bougeault and Lacarrère 1989); the University of Washington (UW) scheme (Bretherton and Park 2009), Grenier–Bretherton–McCaa (GBM; Grenier and Bretherton 2001); Shing–Hong (SH) scheme (Shin and Hong 2015); and MYNN, level 3 (MYNN3; Nakanishi and Niino 2006). Further information about these schemes can be found in Banks et al. (2016) and Cohen et al. (2015).

WRF is run in a nested configuration with the domains shown in Fig. 1a. The outermost grid, at a horizontal resolution of 15 km, comprises the entire Scandinavian Peninsula. The 3-km grid includes the full Botnia–Atlantica region that consists of the Nordic counties of Nordland, Norway; Västerbotten, Sweden; and Pohjanmaa, Finland. The cumulus scheme is switched off in the second domain, while in the outermost grid analysis, nudging toward CFSR is employed. The nudging regions are given in Fig. 1b: the water vapor mixing ratio is nudged in the mid- and upper troposphere, whereas the horizontal wind components and potential temperature perturbation are nudged in the upper troposphere and lower stratosphere, with the nudging time scale set to 1 h for all variables. To prevent the reflection of waves near the model top, a sponge layer is included (Skamarock et al. 2008). Figure 1b also shows the model vertical levels: 60 levels, concentrated in the PBL with about 20 levels in the lowest 200 m. The model output of domain 1 is discarded, and that of domain 2 is stored every 1 h and used for analysis.

3. Statistical model

The BHM provides a framework to characterize spatial or spatiotemporal data (Cressie and Wikle 2011). It is a statistical model consisting of multiple layers, typically three. The top layer, denoted as data model, is devoted to the data vector $\mathbf{y}$ modeled by a certain distribution $[\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\Psi}]$, conditional on the latent variables (defined as those that are not directly measured but rather inferred from observed variables) $\boldsymbol{\theta}$ and the parameters $\boldsymbol{\Psi}$. In the second layer, often called latent or

FIG. 1. (a) Spatial extent of the 15- (blue) and 3-km (red) WRF grids and (b) model (eta) levels with the nudging regions and sponge layer highlighted in red and green rectangles, respectively. The vertical coordinate in (b) is in logarithmic scale.

process model, the latent variables are modeled by the distribution $[\theta|\Psi]$, conditional on the parameters Ψ , in order to characterize the complex spatiotemporal dependencies in the data that are unobservable. In the last layer, the prior distributions for the parameters $[\Psi]$, are provided. Through the product of the three (conditional) distributions, as shown in the equation below, one can obtain the joint distribution and make inference on θ and Ψ based on $\mathbf{y}$, $[\theta, \Psi|\mathbf{y}]$:

$$[\theta, \Psi|\mathbf{y}] \propto [\mathbf{y}|\theta, \Psi][\theta|\Psi][\Psi]. \quad (4)$$

Let the data vector $\mathbf{y} = \{y_{s,t}; i = 1, 2, \dots, N; t = 1, 2, \dots, T\}$, where $y_{s,t}$ denotes the daily mean 2-m temperature as predicted by WRF at a given location s_i and time t . The top layer is modeled as follows:

$$\text{Layer 1: } y_{s,t} | \{\mu_{s,t}, \sigma_e\} = \mu_{s,t} + \epsilon, \quad (5)$$

where $\theta = \{\mu_{s,t}; i = 1, 2, \dots, N; t = 1, 2, \dots, T\}$ is the mean of $\mathbf{y}$, ϵ is a Gaussian noise with mean 0 and variance of σ_e^2 accounting for measurement error, and $\sigma_e \in \Psi$. The $y_{s,t}$ are assumed to be conditionally independent. In the second layer, the latent variables θ are modeled as

$$\text{Layer 2: } \mu_{s,t} | \{\beta_0, \beta'_m s, \xi_{s,t}\} = \beta_0 + \sum_{m=1}^M \beta_m x_{ms_i} + \xi_{s,t}, \quad (6)$$

where $\{\beta_0, \beta'_m s\} \subset \Psi$, β_0 is the intercept term, x_{ms_i} is the m th covariate, β_m is the m th coefficient of x_{ms_i} , and M is the number of covariates. The first two terms on the right-hand side of Eq. (6) express the mean $\mu_{s,t}$ as a multiple linear regression, while $\xi_{s,t}$ is a noise term that accounts for the spatiotemporal effect; $\xi_{s,t}$ is modeled through its first-order neighbor in time and neighbors in space as

$$\xi_{s,t} = \rho \xi_{s,(t-1)} + \phi_{s,t}, \quad (7)$$

where $t = 1, 2, \dots, T$. In Eq. (7), ρ characterizes the temporal correlation between $\xi_{s,t}$ and $\xi_{s,(t-1)}$, and $\phi_{s,t}$ represents the spatial effect specified by a Gaussian Markov random field (GMRF). A GMRF is a random vector following a multivariate Gaussian distribution. Any element of the random vector, given its neighbors, is conditionally independent of all the other elements. A GMRF is used here because (i) usually the data in nature follows a Gaussian distribution, (ii) it is generally a good starting point for statistical methods, and (iii) it approximates the Bayesian inference with high

accuracy. Let $|\rho| < 1$ and $\xi_{s,1} \sim N[0, \sigma_M^2/(1 - \rho^2)]$ where σ_M^2 is the marginal variance of $\phi_{s,t}$, such that $\xi_{s,t}$ is a weakly stationary process in time. As shown in Lindgren et al. (2011), one can represent $\phi_{s,t}$, a continuous indexed Gaussian field with Matérn covariance (Matérn 1960), as a discrete indexed GMRF with Matérn covariance defined in the triangulation over the spatial domain (i.e., the domain is subdivided into spatial elements that have the minimum number of

boundary points; for a two-dimensional space, the domain is subdivided into triangles). Note that the larger the number of vertices in the triangulation, the higher the accuracy of the GMRF representing it, which comes at an increased computational cost. Blangiardo et al. (2013) illustrates how to configure the triangulation in detail; $\phi_{s,t}$ can be used to represent a discrete indexed GMRF with zero mean and covariance defined by

$$\text{Cov}(\phi_{s,t_1}, \phi_{s,t_2}) = \begin{cases} \text{Matérn covariance,} & \text{if } t_1 = t_2 \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

for $i \neq j$.

Matérn covariance at time point t is defined as

$$\begin{aligned} \text{Cov}(\phi_{s,t_1}, \phi_{s,t_2}) &= \frac{\sigma_M^2}{\Gamma(\nu) \times 2^{\nu-1}} (\kappa \|s_i - s_j\|_2)^\nu \\ &\times K_\nu(\kappa \|s_i - s_j\|_2), \end{aligned} \quad (9)$$

where K_ν is the modified Bessel function of the second kind, $\Gamma(\cdot)$ is the Gamma function, ν is a smoothness parameter, and κ is a range parameter. In Eq. (9), $\|s_i - s_j\|_2$ represents the Euclidian (i.e., straight line) distance between the sites s_i and s_j . Two important properties of a random field with Matérn covariance are (i) it is stationary and isotropic (when the distance is Euclidean), and (ii) the covariance decreases as the distance between s_i and s_j increases [note that $K_\nu(X)$ very rapidly approaches zero with increasing X]. This second property of a Matérn covariance appropriately reflects the relationship of the temperature over the spatial domain in our case and justifies the use of this particular covariance in this work. The ratio $\sqrt{8\nu}/\kappa$ represents the distance at which the covariance is about 0.1 (Matérn 1960; Stein 1999). If two points are spaced farther apart than this distance, the corresponding random variables are weakly correlated or uncorrelated.

The parameters ρ , σ_M , κ , ν used in Eqs. (7) and (9) specify the latent spatiotemporal process and are included in the parameter space Ψ . In the last layer, some priors are provided for all these parameters:

Layer 3: Provides prior distributions for the parameters, i.e., Ψ .

The posterior mean of a latent variable/parameter is used as its Bayes estimator: for example, $E(\mu_{s,t}|\mathbf{y})$ is the Bayes estimator of $\mu_{s,t}$. Posterior means of all the latent variables/parameters are derived from the joint posterior distribution following Eq. (4).

The INLA (Rue et al. 2009) package in the software R is used to implement the BHM as described above. The R code is quite straightforward; it codes the BHM model as presented above directly into the R language, and is freely available online.¹

4. Methods

In this section, the setup of the BHM and the diagnostics used to evaluate the performance of the WRF-BHM are presented.

a. WRF data

The focus is on the daily mean 2-m temperature from the 3-km WRF grid, Fig. 1a, which contains 301×381 grid points. The BHM is set up in a 20-day period in the winter (1–20 January 2017) and summer (1–20 July 2016) seasons. For each month, the BHM was first fit to the WRF data using different length periods multiple of 5 days and up to 30 days. It was found that there was very little sensitivity in the parameters for a period longer than 20 days. Hence, and for computational reasons, a 20-day period for each season is selected to develop the BHM. The variance σ_e^2 is assumed to be the same for all time and space points. While it may be reasonable to consider that at a given grid point, σ_e^2 is constant in time (at least for a given month), spatial homogeneity is unlikely: for example, a coastal location is expected to have a smaller temperature variance than an inland site, with the latter more likely to experience extreme values. Given this, possible spatial heterogeneity in σ_e^2 is investigated in this study. Four approaches are considered:

- Method I: The 10000 grid points ($N = 10000$) are randomly selected from the available 301×381 in the

¹ <http://www.r-inla.org/examples/tutorials/spde-tutorial>.

model domain. The assumption is that the variance σ_e^2 is spatially homogeneous in the domain considered.

- Method II: The 10 000 grid points ($N = 10\,000$) around coastlines and inland sites randomly are selected. A grid point is considered to be a coastline point if one of its nearest spatial neighbors is totally (or partially) covered by an ocean or sea. It is considered to be an inland point if its nearest coastal point is beyond its third-order neighbors. Here, it is assumed that spatial heterogeneity in σ_e^2 arises mostly from the proximity to water bodies, in particular to the Atlantic Ocean and Gulf of Bothnia.
- Method III: The 10 000 grid points ($N = 10\,000$) at two different elevations (namely, ≤ 673.2 and > 673.2 m) are randomly selected. A threshold of 673.2 m is chosen as to have at least 10 000 grid points in each category. The assumption is that spatial heterogeneity in σ_e^2 arises mostly from the elevation of the site.
- Method IV: The domain is divided into nine regions (shown in Fig. A1) each containing a minimum of 10 000 grid points ($N = 10\,000$). The purpose of this approach is to have the largest possible spatial heterogeneity in σ_e^2 with the available data.

In a nutshell, a total of 10 000 grid points $\times$ 20 days' worth of data from the model simulations are used to set up the BHM for each season.

b. BHM parameter setting

1) COVARIATES

While the 2-m temperature is a function of many covariates, for simplicity only one, the elevation, is considered (i.e., $M = 1$). Future extensions of the BHM developed here may consider additional covariates such as the latitude, land-use type, and the proximity to a water body.

2) PRIOR DISTRIBUTIONS FOR PARAMETERS

How to build the prior distribution (informative or noninformative approach) of a given parameter is a challenging task in Bayesian analysis and will depend on whether there is preexisting information about the parameter being modeled. Both approaches have their strengths and weaknesses (e.g., Berger 2006). When one has limited information about the parameter, weakly informative prior is generally a good choice (Simpson et al. 2015). In this study, the default specifications in INLA for the distributions of the parameters are used, which correspond to noninformative prior for β_0 and weakly informative priors for the others.

The noninformative prior assigned to β_0 is Gaussian distributed with a mean of 0 and a precision (inverse of

variance) of 0, while the β_m values follow a Gaussian distribution with a mean of 0 and a precision of 0.001. The $\log(1/\sigma_e^2)$ is set to $\log\Gamma(1.5 \times 10^{-5})$, which is defined as a random variable $X \sim \log\Gamma(a, b)$ if $\exp(X) \sim \Gamma(a, b)$ (Martino and Rue 2010); $\log[(1 + \rho)/(1 - \rho)]$ is assumed to follow a normal distribution with a mean of 0 and a precision of 0.15; $\log(\kappa)$ and $\log(1/\sigma_M^2)$ are assumed as $N(0, 1)$. The variable ν is set to 1, which corresponds to a third-order conditional autoregressive structure under regular lattice data. Although other values for ν are possible, they have not been fully tested in the INLA package and hence are not considered.

c. BHM evaluation

Four diagnostics are employed with the aim of assessing different aspects of the WRF-BHM. Diagnostics D1 and D2 are computed at every grid point using the WRF-BHM and the 10 day $\times$ 10 day summer and winter experiments and compare the mean and standard deviation obtained from the two methods using the root-mean-square error (RMSE). They are defined in Eqs. (10) and (11) below:

$$\text{D1: RMSE}_{s_i} = \sqrt{\frac{\sum_{t=1}^T [(\hat{\mu}_e)_{s_{it}} - \hat{\mu}_{s_{it}}]^2}{T}}, \quad (10)$$

where $(\hat{\mu}_e)_{s_{it}}$ is the WRF-BHM estimate of the mean 2-m temperature at the i th grid point and time t , while $\hat{\mu}_{s_{it}}$ is the sample mean calculated from the 10 WRF simulations with $T = 10$ days. Diagnostic D2 is as D1, but for the standard deviation:

$$\text{D2: RMSE}_{s_i} = \sqrt{\frac{\sum_{t=1}^T (\hat{\sigma}_e - \hat{\sigma}_{s_{it}})^2}{T}}, \quad (11)$$

where $\hat{\sigma}_e$ is the WRF-BHM estimate of the standard deviation for a given method, and $\hat{\sigma}_{s_{it}}$ is the sample standard deviation at the i th grid point and time t calculated from the 10 WRF simulations (also $T = 10$ days).

The next two diagnostics assess how close the temperature distributions estimated from WRF-BHM and the traditional multiphysics ensemble experiment are. The WRF-BHM temperature distribution is Gaussian by assumption, Eq. (5), while lacking additional information, the temperature distribution from the multiphysics ensemble at a given day and grid point is obtained using the kernel density estimation method with the 10 ensemble members. The third indicator is the Kolmogorov–Smirnov (K-S) statistic (Massey 1951) given by

TABLE 1. Summary statistics (including mean μ , standard deviation σ , and quantiles Q) of the posterior distributions of the BMH parameters for the winter experiment. For method IV, only the minimum and maximum values are given, full data is provided in Table A1.

Group	Parameter	μ	σ	$Q_{2.5\%}$	$Q_{97.5\%}$
Method I	β_1 (K m ⁻¹)	-0.0067	0.000	-0.0068	-0.0067
	σ_e (K)	0.609	0.001	0.607	0.611
	ρ	0.832	0.006	0.820	0.846
	$\sqrt{8\nu/\kappa}$ (km)	11.218	0.545	10.178	12.322
	β_1 (K m ⁻¹)	-0.0063 (-0.0049)	0.000 (0.000)	-0.0063 (-0.005)	-0.0063 (-0.0049)
Method II: coastal (inland)	σ_e (K)	0.480 (0.536)	0.001 (0.001)	0.478 (0.534)	0.481 (0.538)
	ρ	0.738 (0.644)	0.007 (0.008)	0.722 (0.627)	0.751 (0.658)
	$\sqrt{8\nu/\kappa}$ (km)	9.062 (10.764)	0.425 (0.671)	8.269 (9.349)	9.940 (11.942)
	β_1 (K m ⁻¹)	-0.0078 (-0.0056)	0.000 (0.000)	-0.0076 (-0.0056)	-0.0078 (-0.0056)
	σ_e (K)	0.591 (0.670)	0.001 (0.001)	0.589 (0.668)	0.593 (0.672)
Method III: low elevation (high elevation)	ρ	0.846 (0.731)	0.004 (0.018)	0.835 (0.693)	0.854 (0.767)
	$\sqrt{8\nu/\kappa}$ (km)	9.944 (7.406)	0.831 (0.648)	8.124 (6.485)	11.261 (8.958)
	β_1 (K m ⁻¹)	-0.0146 (-0.0026)	0.000 (0.000)	-0.0147 (-0.0027)	-0.0145 (-0.0025)
	σ_e (K)	0.086 (0.744)	0.001 (0.001)	0.085 (0.742)	0.086 (0.747)
	ρ	0.619 (0.980)	0.001 (0.010)	0.602 (0.976)	0.640 (0.982)
Method IV: min (max)	$\sqrt{8\nu/\kappa}$ (km)	4.365 (10.698)	0.202 (0.813)	4.011 (9.303)	4.847 (12.484)

$$D3: D_{KS_{s_i}} = \max_t \left\{ \sup_x \left[|F_{W_{s_{it}}}(x) - F_{E_{s_{it}}}(x)| \right] \right\}, \quad \text{for} \\ t = 1, 2, \dots, T; \quad x = 1, 2, \dots, X, \quad (12)$$

where $F_{W_{s_{it}}}$ and $F_{E_{s_{it}}}$ represent the cumulative distribution functions (CDFs) of the WRF-BHM and the ensemble 2-m temperature data at the i th grid point and time t , respectively. At a given grid point and for a given day, the two CDFs are sampled $X = 1000$ times with the maximum distance between them being stored. This procedure is then repeated for the other days with the diagnostic returning the maximum distance over $T = 10$ days. The calculations were repeated for different values of X , and it was concluded that the sensitivity of the K-S statistic to X is almost negligible, at least for X values in the range 500–10 000 (not shown). The optimal score is 0, in which case the two CDFs overlap each other for the full period, and the worst score is 1.

The fourth diagnostic is the Kullback–Leibler (K-L) divergence (MacKay 2003) defined as

$$D4: D_{KL_{s_i}} = \frac{1}{T} \sum_{t=1}^T \left\{ \sum_{x=1}^X P_{E_{s_{it}}}(x) \log \left[\frac{P_{E_{s_{it}}}(x)}{P_{W_{s_{it}}}(x)} \right] \right\}, \quad (13)$$

where $P_{W_{s_{it}}}(x)$ and $P_{E_{s_{it}}}(x)$ are the probability distribution functions (PDFs) of WRF-BHM and ensemble 2-m temperature data at the i th grid point and time t , respectively. This diagnostic is also computed over $T = 10$ days. For a given day, P_W and P_E are sampled 1000 times (i.e., $X = 1000$) in the range [268, 300 K] in the summer period and [238, 290 K] in the winter period to compute the K-L divergence for that day. The K-L divergence scores are found to be not insensitive to the

number of samples X , at least for X values in the interval 500–10 000 (not shown). The ranges of temperatures selected for each season for sampling the two PDFs cover the full range of temperatures obtained with the WRF-BHM and ensemble approaches. The K-L divergence values for each of the 10 days are then averaged to generate the D4 diagnostic. The K-L divergence can be interpreted as a nonsymmetric (i.e., not invariant to a change in the order of P_W and P_E) measure of the difference between the two PDFs at a given model grid point. As with D3, the optimal score is 0 obtained when the two PDFs overlap, where the closer to zero, the better the agreement between the predicted and observed temperature distributions.

5. Results and discussion

a. Winter experiment (January 2017)

Table 1 shows the summary statistics of the posterior distributions of the BHM parameters for this season. Only parameters statistically significantly different from zero at the 95% confidence interval [$Q_{2.5\%}$, $Q_{97.5\%}$], where Q represents the quantile, are shown. In addition, for method IV, only the minimum and maximum of the summary statistics over the nine regions are given; full results are provided in the appendix. The marginal variance σ_M^2 , which explains partially the variability of $\mu_{s_{it}}$, is not reported as it is not of interest.

Regarding method I, β_1 is constant across the domain implying, from Eq. (6), a linear decrease of the 2-m temperature with elevation at a rate of -6.7 K km^{-1} . This value is close to the average environmental lapse rate of -6.5 K km^{-1} . The parameter σ_e follows a

FIG. 2. Diagnostic D1, an indication of how similar the WRF-BHM's mean 2-m temperature is to that estimated from an ensemble of WRF simulations with different PBL schemes, for the (a) winter period (1–10 Jan 2017) and (b) summer period (1–10 Jun 2016). The optimal value for diagnostic D1 is zero.

distribution with a mean of 0.609 K and a standard deviation of 0.001 K, indicating that the uncertainty in the WRF-predicted 2-m temperature is typically of about 0.61 K. The correlation ρ used in $\xi_{s,t}$ defined in Eq. (7) is rather high, suggesting that the temperature variability at a given grid point is strongly linked to that of its closest neighbors. This gives further credibility to the BHM approach followed here, which is built on the premise that the closest neighbors of a given grid point can provide useful information about the state of a field at that grid point. The same conclusion is reached by looking at the last parameter shown, $\sqrt{8\nu/\kappa}$, which indicates the distance between two points beyond which the corresponding temperatures are weakly correlated or uncorrelated. For method I, this distance is typically 10–12 km, which corresponds to 3–4 model grid points, and therefore includes the closest spatial neighbors.

Comparable values are obtained for methods II–IV. For method II, a lower σ_e for coastal sites is expected as they are likely to have a reduced temperature variability when compared to inland regions. For method III, the higher-magnitude β_1 for the low-terrain category (≤ 673.2 m) is indicative of steeper lapse rates, closer to the dry adiabatic lapse rate (~ 9.8 K km $^{-1}$). This is consistent with the expected strong thermal inversions that develop here as a result of cold-air pooling (e.g., Pepin et al. 2009). The increased σ_e for the high-elevation category (> 673.2 m) reflects a wider temperature distribution with a more frequent occurrence of extremes. For method IV, a wide range of values for all parameters is obtained, highlighting a strong spatial variability in the domain considered. In line with the findings of methods II and III (as shown in Table A1 in the appendix), the highest σ_e values are found in the northern and eastern parts of the domain, closer to the Arctic and featuring more mountainous terrain, with lower σ_e values in the southern part of the grid, which is relatively flat and where the steepest lapse rates are found. For all methods, ρ is generally reduced when σ_e is larger, indicating an increasing difficulty of the model to capture

the 2-m temperature distribution if the latter is broader and contains more extreme values.

Having fully set up the BHM, the results for the four diagnostics will now be presented. Figure 2a shows the diagnostic D1 for the winter period. Over water bodies and southwestern Finland, the RMSE values are rather small, generally less than 0.5 K. Conversely, and mostly over the northeastern part of the domain in northern Sweden and Finland, they are generally in excess of 0.75 K. Here, rather cold surface temperatures associated with stably stratified boundary layers are common in winter (e.g., Pepin et al. 2009). The WRF Model is known to underperform in such environments (e.g., García-Díez et al. 2013) as some of the dominant physical processes are not well modeled (e.g., Steeneveld 2014). By using different PBL schemes that represent the boundary layer dynamics in different ways (e.g., Cohen et al. 2015; Banks et al. 2016), a wide range of values for the 2-m temperature is obtained, resulting in a larger discrepancy between the two approaches. An inspection of the temperature data revealed that for a given day in the 10-member ensemble experiment, a temperature range in excess of 10 K is obtained in the Arctic Scandinavia, whereas farther south the range is generally smaller, typically less than 2 K (not shown). Larger RMSE values are also seen in valleys and low-elevation regions, where cold-air pooling is a common occurrence in winter, and over the high terrain, where the WRF surface-temperature forecasts are known to be sensitive to the choice of the PBL scheme (e.g., Zhang et al. 2013). Elsewhere, the RMSE values are mostly in the range 0.5–0.75 K. The results in Fig. 2a show that the mean 2-m temperature estimates using the two approaches are generally similar (RMSE differences mostly within 0.75 K). Figure 3 shows the results for σ_e for the four methods. The spatial pattern of the RMSE values of the standard deviation resembles that of the RMSE of the mean shown in Fig. 2a with higher values in the northeastern part of the domain and lower values elsewhere, in particular in southern and

FIG. 3. Diagnostic D2, an indication of how similar the WRF-BMH's σ_e is to the uncertainty estimated from an ensemble of WRF simulations with different PBL schemes, obtained from (a) method I, (b) method II, (c) method III, and (d) method IV for the period 1–10 Jan 2017. The optimal value for diagnostic D2 is zero.

southwestern parts. According to Table 1, σ_e is typically of 0.6 K, and hence, the RMSE values obtained here are about a third to half of that magnitude. This indicates that the uncertainty estimated from just one WRF experiment using the BHM approach is comparable to that obtained from an ensemble of simulations with perturbed model physics. This is confirmed by looking at the uncertainties themselves (not shown). As no method is found to clearly outperform another, method I will be considered in the subsequent analysis, given its simplicity.

To illustrate the closeness between the WRF-BHM and ensemble temperature distributions, Figs. 4a and 4c show diagnostics D3, the K-S statistic, and D4, the K-L divergence, respectively. Both metrics are nonnegative with the optimal value being zero. The K-S values are mostly within 0.8, being less than 0.5 over the water bodies and in parts of southern Sweden and southern and western Finland. Here, the WRF-BHM and ensemble 2-m temperature CDFs are expected to be more similar given the smaller RMSE values shown in Figs. 2a and 3. Diagnostic D4 gives a similar picture with the two PDFs being more divergent in the northeastern part of the domain where their means and standard deviations are farther apart. Quantitatively speaking, the scores for the K-S and K-L diagnostics shown in Fig. 4 are not very good. As shown in Fig. 5, in which an explicit comparison of the two temperature distributions at several locations in the model domain is performed, and

as discussed in more detail below, the lower K-L (and K-S) scores arise from a more significant deviation of the ensemble temperature distribution from a Gaussian shape.

As to allow for a more direct evaluation in the left panels of Fig. 5 and at a given day and for two grid points over land and water for which higher and lower D1, D2, and D4 (similar results for D3) scores are obtained, the WRF-BHM and ensemble temperature distributions for the winter season are plotted. It is important to note that similar results are obtained at other grid points (not shown), and hence these curves are representative of those obtained elsewhere. Regarding the shape of the distribution, and despite the presence of a secondary maximum, the ensemble temperature distribution generally resembles a Gaussian distribution, supporting the assumption of a Gaussian shape made in section 3 for the BHM. What is more, and as expected, the agreement between the curves is better if the D1, D2, and D4 scores are higher, which also confirms that the use of these verification diagnostics for this work is justified. For all the winter results, the mean value of the WRF-BHM temperature distribution is within 1 K of that obtained with the ensemble of simulations. Because of the presence of the referred secondary maximum, possibly arising from the small number (10) of ensemble members considered, there are some differences in the spread of the two distributions. In any case, and when taken as a whole, the results given in Fig. 5 highlight that the

FIG. 4. Diagnostic D3, K-S statistic, for the (a) winter period (1–10 Jan 2017) and (b) summer period (1–10 Jun 2016). (c), (d) As in (a) and (b), but for the diagnostic D4, K-L divergence. Diagnostics D3 and D4 are nonnegative with the optimal value being zero.

simpler and computationally cheaper WRF-BHM approach can be used to estimate the ensemble-generated temperature distribution at least for this region and season.

b. Summer experiment (July 2016)

Table 2 is as Table 1, but for the summer experiment. For this season, β_1 is in the range from -6 to -8 K km^{-1} , typical values of the environmental lapse rate, and σ_e is mostly between 0.3 and 0.5 K . A comparison with the corresponding winter values reveals a smaller σ_e for the summer period. During the warm season, the occurrence of cold and stably stratified environments is infrequent with the longer daylight leading to a more convectively active boundary layer and stronger vertical mixing. This is likely to lead to a smaller variability of the 2-m temperature, which is reflected in the reduced values of σ_e and in the higher values of $\sqrt{8\nu/\kappa}$. These results have an important implication: temporal homogeneity cannot be assumed at least for some of the BHM parameters for a full year. Should this technique be extended to comprise a complete annual cycle, in a high-latitude region with complex topography such as Botnia–Atlantica, the BHM would have to be fitted to each season (or perhaps each month) separately. Nevertheless, the computational resources needed to do so are much less than those required to perform an ensemble of simulations.

The correlation ρ obtained with WRF data is still high in excess of 0.5 but smaller than the winter one by up to 30% . The lower correlation values seem to come mostly from the coastal regions where strong thermal circulations that normally develop in this season (e.g., Smedman et al. 1995) may not be well represented by the model. The higher σ_e for high-elevation sites noted in the winter experiment in method III is also present, while the steeper lapse rates β_1 obtained for the low-elevation regions in the cold season are no longer seen as cold-air pooling is not significant in the summer. The results for method IV (given in Table A2) show generally smaller differences in the parameter values for the nine regions compared to the winter experiment. As is the case for the cold season, ρ is lower when σ_e is higher, reflecting an increasing difficulty of the BHM to simulate the observed temperature variability when the latter is larger.

The results for diagnostics D1 and D2 for the summer period are given in Figs. 2b and 6, respectively. The RMSE for the mean temperature is smaller than that for the winter season with values generally less than 0.5 K everywhere in the domain. RMSE values in excess of 0.5 K are seen mostly in coastal regions. As stated before, land–sea breeze circulations and associated temperature variations are a common occurrence in the summer along coastlines, and their representation is sensitive to the PBL option chosen

FIG. 5. The 2-m temperature distribution at four grid points for which the diagnostics D1, D2, and D4 indicate good and not-so-good WRF-BHM performance for the first day of the winter (1 Jan 2017) and summer (1 Jul 2016) periods. The blue curve is the WRF-BHM temperature distribution, and the red curve is that obtained from the multiphysics ensemble experiment. For a given grid point, the ensemble temperature distribution is calculated using the kernel density estimation method with the 10 ensemble members. See the title of each plot for information regarding the location of the grid points and the scores.

TABLE 2. Summary statistics (including mean μ , standard deviation σ , and quantiles Q) of the posterior distributions of the BMH parameters for the summer experiment. For method IV, only the minimum and maximum values are given, full data is provided in Table A2.

Group	Parameter	μ	σ	$Q_{2.5\%}$	$Q_{97.5\%}$
Method I	β_1 (K m^{-1})	-0.0068	0.000	-0.0068	-0.0068
	σ_e (K)	0.331	0.001	0.330	0.333
	ρ	0.630	0.007	0.617	0.647
	$\sqrt{8\nu/\kappa}$ (km)	15.051	0.727	13.536	16.369
Method II: coastal (inland)	β_1 (K m^{-1})	-0.0070 (-0.0074)	0.000 (0.000)	-0.0070 (-0.0074)	-0.0070 (-0.0074)
	σ_e (K)	0.288 (0.263)	0.000 (0.000)	0.287 (0.262)	0.289 (0.264)
	ρ	0.508 (0.632)	0.013 (0.010)	0.479 (0.612)	0.530 (0.652)
	$\sqrt{8\nu/\kappa}$ (km)	15.317 (14.744)	1.067 (0.721)	13.041 (13.391)	17.145 (16.227)
Method III: low elevation (high elevation)	β_1 (K m^{-1})	-0.0060 (-0.0067)	0.000 (0.000)	-0.0061 (-0.0067)	-0.0060 (-0.0077)
	σ_e (K)	0.323 (0.392)	0.001 (0.001)	0.321 (0.391)	0.324 (0.393)
	ρ	0.631 (0.667)	0.002 (0.019)	0.626 (0.623)	0.637 (0.701)
	$\sqrt{8\nu/\kappa}$ (km)	14.961 (12.199)	0.136 (0.692)	14.664 (11.029)	15.191 (13.728)
Method IV: min (max)	β_1 (K m^{-1})	-0.0073 (-0.0045)	0.000 (0.000)	-0.0073 (-0.0046)	-0.0073 (-0.0044)
	σ_e (K)	0.076 (0.475)	0.000 (0.001)	0.075 (0.473)	0.076 (0.477)
	ρ	0.351 (0.959)	0.002 (0.018)	0.325 (0.956)	0.380 (0.964)
	$\sqrt{8\nu/\kappa}$ (km)	4.230 (12.820)	0.200 (0.963)	3.820 (11.386)	4.600 (14.531)

(e.g., Steele et al. 2013, 2015). Hence, and at these sites, a larger difference between the predicted 2-m mean temperatures using the two methods is likely to occur. An inspection of the temperatures for the ensemble experiment reveals that in coastal regions the predicted 2-m temperatures can differ by as much as 3 K, whereas at inland sites they are mostly within 1 K (not shown). Larger RMSE values are also seen over the high terrain where the model temperature forecasts are known to be sensitive to the PBL scheme used (e.g., Zhang et al.

2013). As is the case for the diagnostic D1, the RMSE of the standard deviation estimates, given in Fig. 6, is also generally smaller than that obtained in the winter season. As for the summer period, σ_e is typically of 0.3–0.5 K (Table 2) the RMSE values are about half of that magnitude, and hence, the uncertainty estimated using the WRF-BHM is comparable to that obtained from the ensemble of simulations. The only regions for which diagnostic D2 exceeds 0.4 K are a portion of central-eastern Sweden and parts of Finland. Here, there is a

FIG. 6. As in Fig. 3, but for the summer period (1–10 Jul 2016).

larger spread in the temperatures predicted by the different members of the ensemble (not shown). It is possible that more regional and local-scale circulations develop, leading to an increased sensitivity to the choice of the PBL scheme. A further investigation into why the variability is larger in these regions is beyond the scope of this paper. Similar scores are obtained regardless of the assumptions made on the spatial heterogeneity of σ_e . As for the winter season, method I is selected for further analysis given its simplicity.

Figures 4b and 4d show the K-S and K-L divergence diagnostics for the summer season. The magnitude and spatial pattern of the two scores are similar to those obtained in the winter season with lower values mostly over the Gulf of Bothnia and Atlantic Ocean and higher values in some inland sites. As opposed to the winter period, the lowest scores are not seen in the northeastern (Arctic) part of the domain but in some coastal and high-terrain regions. Here, as shown in Figs. 2b and 6, the mean and standard deviation of the 2-m temperature distributions obtained using the WRF-BHM and multiphysics ensemble differ the most, so the correspondent CDFs and PDFs are expected to be farther apart. These lower scores can be attributed to a more significant deviation of the ensemble temperature distribution from a Gaussian shape as seen in Fig. 5.

The right panels in Fig. 5 show the two temperature distributions at four grid points for the summer season (similar results are obtained at other locations in the domain, not shown). As for the winter case, the WRF-BHM distribution generally resembles that obtained with an ensemble of simulations with the respective means being within 1 K of each other. The presence of a secondary maximum in the ensemble temperature distribution may be an artifact of the limited number of samples used. The results presented here highlight the success of the proposed WRF-BHM formulation in simulating the daily mean 2-m temperature distribution as given by the multiphysics ensemble.

6. Summary and conclusions

When running a numerical model, the gridbox-averaged value is a standard output, but the uncertainty that goes with that prediction is generally not provided. A common way to estimate the uncertainty associated with weather and climate projections is to perform an ensemble of model simulations. This can be achieved in different ways, for example, by changing the model physics and/or dynamics configuration, perturbing the initial and/or boundary conditions, or even by considering different models. In this paper, we propose a novel approach that only requires one simulation:

fitting a BMH to the model data and from the estimated distribution extracting the uncertainty of a given variable (here, the 2-m temperature, a crucial field for climate projections, is considered). It is important to note that WRF-BHM is not a true alternative to ensemble forecasting in the sense that it cannot generate the same set of metrics. For example, the most likely event (e.g., for precipitation type) cannot be obtained with the BHM. However, and for commonly used ensemble products such as the uncertainty estimation, probability of exceedance, and background-error covariance for ensemble Kalman filter and hybrid variational data assimilation schemes, it provides an easy to apply, computationally cheap, and faster alternative to ensemble forecasting. We would also like to stress that, although not discussed here, the BHM can also be run in prediction mode. In particular, it can analyze data possessing spatiotemporal dependence and make use of the estimated dependence to predict the values of a given field at different locations and time periods. However, it can be argued that using the BHM for an understanding of the spatiotemporal dependency inherent to a dataset is more relevant than using it for prediction purposes.

WRF is run at 3-km spatial resolution over the Botnia–Atlantica region for one month in the winter (January 2017) and summer (July 2016) seasons. In addition, an ensemble of simulations is generated with 10 different PBL schemes for the first 10 days of January 2017 and July 2016. The uncertainty sampled in this work is that associated with the way the boundary layer dynamics and near-surface flow are represented in the model and their impact on the daily mean 2-m temperature predictions. A fundamental assumption of comparing the BHM to the multiphysics ensemble is that the error is primarily associated with PBL physics uncertainties and that the multiphysics ensemble provides a meaningful estimate of that.

The BHM uses nearby information in both space and time to estimate statistical properties of our WRF variable of interest at each grid location. The 2-m temperature, at a given time and spatial location, is assumed to follow a normal distribution with mean $\mu_{s,t}$ and standard deviation σ_e . Four assumptions are made to explore spatial heterogeneity in σ_e , ranging from spatial homogeneity to largest possible heterogeneity. Four diagnostics are used to evaluate the performance of the BHM. In the first two, the temperature mean and standard deviation estimated from WRF-BHM is compared to that obtained from the referred ensemble of simulations. The other two diagnostics give information regarding the degree of overlap between the two distributions. For a given day and model grid point, the

FIG. A1. Nine spatial regions used in method IV.

ensemble temperature distribution is calculated using the kernel density estimation method with the 10 ensemble members.

The BHM parameters obtained with WRF data are consistent with the expected values. Given the differences in some of them between the two seasons, and at least for this region, it is not advisable to consider just one BHM for the full year. Studies that wish to extend this work for a complete annual cycle should fit a BHM to at least each season separately, as the regional and local-scale dynamics are very different throughout the year. For both summer and winter seasons, the mean and uncertainty estimated using the WRF-BHM are found to be generally comparable to those obtained from an ensemble of simulations with different PBL schemes. The exceptions are (i) Arctic Scandinavia and low-level regions in winter, where cold and stably stratified boundary layers are ubiquitous; (ii) coastal regions in the summer, where land–sea breeze circulations are more significant; and (iii) over the high terrain in both seasons. Here, the WRF-predicted 2-m temperature has been reported to be particularly sensitive to the choice of the PBL scheme (e.g., [García-Díez et al. 2013](#); [Steele et al. 2015](#); [Zhang et al. 2013](#)) resulting in larger differences in the simulated temperature distributions using the two approaches. A direct inspection of the WRF-BHM and ensemble temperature distributions revealed the presence of a secondary maximum in the latter, with this stronger deviation from a Gaussian shape probably explaining the lower scores. This feature

possibly arises from the limited number of ensemble members considered. Despite differences in the spread, the means of the two distributions are very similar, within 1 K of each other. Hence, the WRF-BHM approach can be used to estimate the uncertainty and, more generally, the distribution of the 2-m temperature as given by the multiphysics ensemble. Since these results are comfortably interpreted based on our general understanding of the weather variables for the Scandinavian Peninsula, we expect the application of the BHM for estimation of modeled weather variable uncertainty to perform similarly in other temporal and spatial domains. What is more, if the ensemble WRF distributions displayed in [Fig. 5](#) were associated with actual surface observations, and we used an appropriate inference method to assess the equality of the distribution pairs, the approach followed in this work then allows a statistical assessment on the equality of the model results and observations. Such inferential statistical analyses have important applications for geospatial research, including climate change.

In conclusion, the main contributions of this work are as follows: (i) we apply the BHM method to estimate the uncertainty of a WRF output variable (2-m temperature) using space- and time-dependent data; (ii) we prove the consistency of the method by demonstrating that the uncertainty estimate correctly summarizes the aggregated space/time data; (iii) we show that the temperature distribution obtained with WRF-BHM is similar to that computed from an ensemble of model

TABLE A1. Summary statistics (including mean μ , standard deviation σ , and quantiles Q) of the posterior distributions of the BMH parameters for each of the nine regions considered in method IV for the winter experiment.

Group	Parameter	μ	σ	$Q_{2.5\%}$	$Q_{97.5\%}$
Region 1	β_1 (K m ⁻¹)	-0.0109	0.000	-0.0109	-0.0109
	σ_e (K)	0.086	0.000	0.085	0.086
	ρ	0.980	0.001	0.976	0.982
	$\sqrt{8\nu/\kappa}$ (km)	10.698	0.813	3.303	12.484
	β_1 (K m ⁻¹)	-0.0072	0.000	-0.0073	-0.0072
Region 2	σ_e (K)	0.744	0.001	0.742	0.747
	ρ	0.844	0.006	0.830	0.856
	$\sqrt{8\nu/\kappa}$ (km)	4.528	0.240	4.033	4.972
	β_1 (K m ⁻¹)	-0.0026	0.000	-0.0027	-0.0025
	σ_e (K)	0.635	0.001	0.633	0.637
Region 3	ρ	0.619	0.009	0.602	0.640
	$\sqrt{8\nu/\kappa}$ (km)	4.494	0.231	4.011	4.913
	β_1 (K m ⁻¹)	-0.074	0.000	-0.0074	-0.0073
	σ_e (K)	0.445	0.001	0.442	0.447
	ρ	0.764	0.007	0.750	0.780
Region 4	$\sqrt{8\nu/\kappa}$ (km)	6.592	0.438	5.669	7.410
	β_1 (K m ⁻¹)	-0.0057	0.000	-0.0057	-0.0056
	σ_e (K)	0.595	0.001	0.593	0.597
	ρ	0.790	0.008	0.774	0.807
	$\sqrt{8\nu/\kappa}$ (km)	4.365	0.218	4.013	4.859
Region 5	β_1 (K m ⁻¹)	-0.0069	0.000	-0.0070	-0.0068
	σ_e (K)	0.324	0.001	0.323	0.325
	ρ	0.776	0.009	0.757	0.795
	$\sqrt{8\nu/\kappa}$ (km)	5.267	0.202	4.908	5.700
	β_1 (K m ⁻¹)	-0.0055	0.000	-0.0055	-0.0055
Region 6	σ_e (K)	0.588	0.000	0.586	0.590
	ρ	0.774	0.008	0.758	0.792
	$\sqrt{8\nu/\kappa}$ (km)	4.402	0.211	4.016	4.847
	β_1 (K m ⁻¹)	-0.0055	0.000	-0.0055	-0.0054
	σ_e (K)	0.389	0.001	0.387	0.390
Region 7	ρ	0.857	0.006	0.844	0.870
	$\sqrt{8\nu/\kappa}$ (km)	6.181	0.312	5.598	6.826
	β_1 (K m ⁻¹)	-0.0146	0.000	-0.0147	-0.0145
	σ_e (K)	0.275	0.000	0.274	0.276
	ρ	0.752	0.010	0.731	0.772
Region 8	$\sqrt{8\nu/\kappa}$ (km)	6.288	0.370	5.671	7.114
	β_1 (K m ⁻¹)	-0.0072	0.000	-0.0073	-0.0072
	σ_e (K)	0.744	0.001	0.742	0.747
	ρ	0.844	0.006	0.830	0.856
	$\sqrt{8\nu/\kappa}$ (km)	4.528	0.240	4.033	4.972
Region 9	β_1 (K m ⁻¹)	-0.0026	0.000	-0.0027	-0.0025
	σ_e (K)	0.635	0.001	0.633	0.637
	ρ	0.619	0.009	0.602	0.640
	$\sqrt{8\nu/\kappa}$ (km)	4.494	0.231	4.011	4.913
	β_1 (K m ⁻¹)	-0.074	0.000	-0.0074	-0.0073

TABLE A2. Summary statistics (including mean μ , standard deviation σ , and quantiles Q) of the posterior distributions of the BMH parameters for each of the nine regions considered in method IV for the summer experiment.

Group	Parameter	μ	σ	$Q_{2.5\%}$	$Q_{97.5\%}$
Region 1	β_1 (K m ⁻¹)	-0.0060	0.000	-0.0060	-0.0060
	σ_e (K)	0.076	0.000	0.075	0.076
	ρ	0.821	0.009	0.801	0.837
	$\sqrt{8\nu/\kappa}$ (km)	10.023	0.530	9.054	11.137
	β_1 (K m ⁻¹)	-0.0066	0.000	-0.0066	-0.0066
Region 2	σ_e (K)	0.475	0.001	0.473	0.477
	ρ	0.752	0.018	0.708	0.778
	$\sqrt{8\nu/\kappa}$ (km)	6.359	0.343	5.754	7.098
	β_1 (K m ⁻¹)	-0.0073	0.000	-0.0073	-0.0072
	σ_e (K)	0.331	0.001	0.330	0.332
Region 3	ρ	0.826	0.008	0.810	0.842
	$\sqrt{8\nu/\kappa}$ (km)	8.807	0.537	7.753	9.860
	β_1 (K m ⁻¹)	-0.069	0.000	-0.0069	-0.0069
	σ_e (K)	0.258	0.009	0.257	0.259
	ρ	0.620	0.015	0.590	0.649
Region 4	$\sqrt{8\nu/\kappa}$ (km)	9.019	0.703	7.549	10.266
	β_1 (K m ⁻¹)	-0.0072	0.000	-0.0072	-0.0072
	σ_e (K)	0.258	0.000	0.257	0.259
	ρ	0.959	0.002	0.956	0.964
	$\sqrt{8\nu/\kappa}$ (km)	4.230	0.200	3.820	4.600
Region 5	β_1 (K m ⁻¹)	-0.0046	0.000	-0.0047	-0.0045
	σ_e (K)	0.221	0.000	0.220	0.222
	ρ	0.694	0.010	0.671	0.711
	$\sqrt{8\nu/\kappa}$ (km)	8.406	0.326	7.833	9.113
	β_1 (K m ⁻¹)	-0.0073	0.000	-0.0073	-0.0073
Region 6	σ_e (K)	0.196	0.000	0.195	0.196
	ρ	0.351	0.014	0.325	0.380
	$\sqrt{8\nu/\kappa}$ (km)	12.820	0.802	11.386	14.531
	β_1 (K m ⁻¹)	-0.0070	0.000	-0.0070	-0.0070
	σ_e (K)	0.181	0.000	0.180	0.181
Region 7	ρ	0.631	0.009	0.614	0.650
	$\sqrt{8\nu/\kappa}$ (km)	11.129	0.622	9.886	12.321
	β_1 (K m ⁻¹)	-0.0045	0.000	-0.0046	-0.0044
	σ_e (K)	0.204	0.000	0.204	0.205
	ρ	0.619	0.013	0.588	0.642
Region 8	$\sqrt{8\nu/\kappa}$ (km)	11.340	0.963	9.875	13.593
	β_1 (K m ⁻¹)	-0.0072	0.000	-0.0073	-0.0072
	σ_e (K)	0.744	0.001	0.742	0.747
	ρ	0.844	0.006	0.830	0.856
	$\sqrt{8\nu/\kappa}$ (km)	4.528	0.240	4.033	4.972
Region 9	β_1 (K m ⁻¹)	-0.0026	0.000	-0.0027	-0.0025
	σ_e (K)	0.635	0.001	0.633	0.637
	ρ	0.619	0.009	0.602	0.640
	$\sqrt{8\nu/\kappa}$ (km)	4.494	0.231	4.011	4.913
	β_1 (K m ⁻¹)	-0.074	0.000	-0.0074	-0.0073

simulations with perturbed physics. The methodology developed here is fully general and can easily be applied to any other variable and numerical model. Possible future extensions of this work include (i) employing more covariates in the process model; (ii) developing a data-driven ν and/or a higher-order autoregressive model in time; (iii) considering a wider domain and/or higher spatial resolution as well as temporal extension in order to have a larger number of pixels that will allow for a better fit of the BHM; (iv) extending the GMRF in space from stationary to nonstationary based, for example, on latitude and land-use type; (v) applying the latent model on σ_e instead of on $\mu_{s,i}$ to reflect heterogeneity within one BHM; and (vi) develop more efficient algorithms to speed up the implementation of the BHM that requires a significant amount of

computer power. In addition, the BHM method can be used in conjunction with an ensemble of simulations to model the distribution of a particular variable. Some of these improvements will be presented in a subsequent paper.

Acknowledgments. We acknowledge Botnia-Atlantica, an EU program financing cross-border cooperation projects in Sweden, Finland, and Norway, for their support of this work through the WindCoE project. We thank the High Performance Computing Center North (HPC2N) for providing the computer resources needed to perform the numerical experiments presented in this paper. We thank two anonymous reviewers for their detailed and insightful comments and suggestions that helped to improve the quality of the paper.

APPENDIX

BHM Parameters for Method IV

In this appendix, the values of the parameters for each of the nine regions (Fig. A1) of method IV are given in Tables A1 and A2 for the winter and summer experiments, respectively.

REFERENCES

- Banks, R. F., J. Tiana-Alsina, J. M. Baldasano, F. Rocadenbosch, A. Papayannis, S. Solomos, and C. G. Tzanis, 2016: Sensitivity of boundary-layer variables to PBL schemes in the WRF Model based on surface meteorological observations, lidar, and radiosondes during the HygrA-CD campaign. *Atmos. Res.*, **176**–**177**, 185–201, <https://doi.org/10.1016/j.atmosres.2016.02.024>.
- Berger, J., 2006: The case for objective Bayesian analysis. *Bayesian Anal.*, **1**, 385–402, <https://doi.org/10.1214/06-BA115>.
- Bitner-Gregersen, E. M., and Coauthors, 2014: Recent developments of ocean environmental description with focus on uncertainties. *Ocean Eng.*, **86**, 26–46, <https://doi.org/10.1016/j.oceaneng.2014.03.002>.
- Blangiardo, M., M. Cameletti, G. Baio, and H. Rue, 2013: Spatial and spatio-temporal models with R-INLA. *Spat. Spatio-Temporal Epidemiol.*, **7**, 39–55, <https://doi.org/10.1016/j.sste.2013.07.003>.
- Bougeault, P., and P. Lacarrère, 1989: Parameterization of orography-induced turbulence in mesobeta-scale model. *Mon. Wea. Rev.*, **117**, 1872–1890, [https://doi.org/10.1175/1520-0493\(1989\)117<1872:POOITI>2.0.CO;2](https://doi.org/10.1175/1520-0493(1989)117<1872:POOITI>2.0.CO;2).
- Bretherton, C. S., and S. Park, 2009: A new moist turbulence parameterization in the community atmosphere model. *J. Climate*, **22**, 3422–3448, <https://doi.org/10.1175/2008JCLI2556.1>.
- Chen, F., and J. Dudhia, 2001: Coupling an advanced land surface–hydrology model with the Penn State–NCAR MM5 modeling system. Part I: Model implementation and sensitivity. *Mon. Wea. Rev.*, **129**, 569–585, [https://doi.org/10.1175/1520-0493\(2001\)129<0569:CAALSH>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2).
- Cohen, A. E., S. M. Cavallo, M. C. Coniglio, and H. E. Brooks, 2015: A review of planetary boundary layer parameterization schemes and their sensitivity in simulating Southeastern U.S. cold season severe weather environments. *Wea. Forecasting*, **30**, 591–612, <https://doi.org/10.1175/WAF-D-14-00105.1>.
- Cressie, N., and C. K. Wikle, 2011: *Statistics for Spatio-Temporal Data*. Wiley, 624 pp.
- DeMaria, M., and J. Kaplan, 1994: A Statistical Hurricane Intensity Prediction Scheme (SHIPS) for the Atlantic basin. *Wea. Forecasting*, **9**, 209–220, [https://doi.org/10.1175/1520-0434\(1994\)009<0209:ASHIPS>2.0.CO;2](https://doi.org/10.1175/1520-0434(1994)009<0209:ASHIPS>2.0.CO;2).
- , M. Mainelli, L. K. Shay, J. A. Knaff, and J. Kaplan, 2005: Further improvements to the Statistical Hurricane Intensity Prediction Scheme (SHIPS). *Wea. Forecasting*, **20**, 531–543, <https://doi.org/10.1175/WAF862.1>.
- De Smedt, T., K. Simons, A. Van Nieuwenhuyse, and G. Molenberghs, 2015: Comparing MCMC and INLA for disease mapping with Bayesian hierarchical models. *Arch. Public Health*, **73** (Suppl.), O2, <https://doi.org/10.1186/2049-3258-73-S1-O2>.
- Dormann, C. F., 2007: Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Ecol. Biogeogr.*, **16**, 129–138, <https://doi.org/10.1111/j.1466-8238.2006.00279.x>.
- Duynkerke, P. G., 1991: Radiation fog: A comparison of model simulations with detailed observations. *Mon. Wea. Rev.*, **119**, 324–341, [https://doi.org/10.1175/1520-0493\(1991\)119<0324:RFACOM>2.0.CO;2](https://doi.org/10.1175/1520-0493(1991)119<0324:RFACOM>2.0.CO;2).
- Evans, J. P., M. Ekström, and F. Ji, 2012: Evaluating the performance of a WRF physics ensemble over South-East Australia. *Climate Dyn.*, **39**, 1241–1258, <https://doi.org/10.1007/s00382-011-1244-5>.
- Fox, N. I., and C. K. Wikle, 2005: A Bayesian quantitative precipitation nowcast scheme. *Wea. Forecasting*, **20**, 264–275, <https://doi.org/10.1175/WAF845.1>.
- García-Díez, M., J. Fernández, L. Fita, and C. Yagüe, 2013: Seasonal dependence of WRF Model biases and sensitivity to PBL schemes over Europe. *Quart. J. Roy. Meteor. Soc.*, **139**, 501–514, <https://doi.org/10.1002/qj.1976>.
- Grenier, H., and C. S. Bretherton, 2001: A moist PBL parameterization for large-scale models and its application to subtropical cloud-topped marine boundary layers. *Mon. Wea. Rev.*, **129**, 357–377, [https://doi.org/10.1175/1520-0493\(2001\)129<0357:AMPPFL>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<0357:AMPPFL>2.0.CO;2).
- Haining, R., J. Law, R. Maheswaran, T. Pearson, and P. Brindley, 2007: Bayesian modelling of environmental risk: A small area ecological study of coronary heart disease mortality in relation to modelled outdoor nitrogen oxide levels. *Stochastic Environ. Res. Risk Assess.*, **21**, 501–509, <https://doi.org/10.1007/s00477-007-0134-1>.
- Hines, K. M., and D. H. Bromwich, 2008: Development and testing of polar WRF. Part I: Greenland Ice Sheet meteorology. *Mon. Wea. Rev.*, **136**, 1971–1989, <https://doi.org/10.1175/2007MWR2112.1>.
- Hong, S.-Y., Y. Noh, and J. Dudhia, 2006: A new vertical diffusion package with an explicit treatment of entrainment processes. *Mon. Wea. Rev.*, **134**, 2318–2341, <https://doi.org/10.1175/MWR3199.1>.
- Iacono, M. J., J. S. Delamere, E. J. Mlawer, M. W. Shephard, S. A. Clough, and W. D. Collins, 2008: Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models. *J. Geophys. Res.*, **113**, D13103, <https://doi.org/10.1029/2008JD009944>.
- Janjić, Z. I., 1994: The step-mountain eta coordinate model: Further developments of the convection, viscous sublayer and turbulence closure schemes. *Mon. Wea. Rev.*, **122**, 927–945, [https://doi.org/10.1175/1520-0493\(1994\)122<0927:TSMECM>2.0.CO;2](https://doi.org/10.1175/1520-0493(1994)122<0927:TSMECM>2.0.CO;2).
- Katata, G., M. Kajino, T. Hiraki, M. Aikawa, T. Kobayashi, and H. Nagai, 2011: A method for simple and accurate estimation of fog deposition in a mountain forest using a meteorological model. *J. Geophys. Res.*, **116**, D20102, <https://doi.org/10.1029/2010JD015552>.
- Koh, T.-Y., and R. Fonseca, 2016: Subgrid-scale cloud-radiation feedback for the Betts-Miller-Janjić convection scheme. *Quart. J. Roy. Meteor. Soc.*, **142**, 989–1006, <https://doi.org/10.1002/qj.2702>.
- Kuligowski, R. J., and A. P. Barros, 1998: Localized precipitation forecasts from a numerical weather prediction model using artificial neural networks. *Wea. Forecasting*, **13**, 1194–1204, [https://doi.org/10.1175/1520-0434\(1998\)013<1194:LPPFAN>2.0.CO;2](https://doi.org/10.1175/1520-0434(1998)013<1194:LPPFAN>2.0.CO;2).
- Leutbecher, M., and T. N. Palmer, 2008: Ensemble forecasting. *J. Comput. Phys.*, **227**, 3515–3539, <https://doi.org/10.1016/j.jcp.2007.02.014>.

- Lindgren, F., H. Rue, and J. Lindström, 2011: An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *J. Roy. Stat. Soc.*, **73B**, 423–498, <https://doi.org/10.1111/j.1467-9868.2011.00777.x>.
- Lo, J. C.-F., Z.-L. Yang, and R. A. Pielke Sr., 2008: Assessment of three dynamical downscaling methods using the Weather Research and Forecasting (WRF) Model. *J. Geophys. Res.*, **113**, D09112, <https://doi.org/10.1029/2007JD009216>.
- MacKay, D. J., 2003: *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 640 pp.
- Malmberg, A., A. Arellano, D. P. Edwards, N. Flyer, D. Nychka, and C. Wikle, 2008: Interpolating fields of carbon monoxide data using a hybrid statistical-physical model. *Ann. Appl. Stat.*, **2**, 1231–1248, <https://doi.org/10.1214/08-AOAS168>.
- Martino, S., and H. Rue, 2010: *Implementing Approximate Bayesian Inference using Integrated Nested Laplace Approximation: A manual for the INLA program*. Department of Mathematical Sciences, Norwegian University of Science and Technology, 72 pp.
- Massey, F. J., Jr., 1951: The Kolmogorov-Smirnov test for goodness of fit. *J. Amer. Stat. Assoc.*, **46**, 68–78, <https://doi.org/10.1080/01621459.1951.10500769>.
- Matérn, B., 1960: Spatial variation: Stochastic models and their application to some problems in forest surveys and other sampling investigations. Ph.D. dissertation, Stockholm University, 144 pp.
- Mills, C. M., 2011: On the Weather Research Forecasting model's treatment of sea ice albedo over the Arctic Ocean. *10th Annual School of Earth, Society, and Environmental Research Review*, Urbana, Illinois, University of Illinois at Urbana-Champaign.
- Monin, A. S., and A. M. Obukhov, 1954: Basic laws of turbulent mixing in the ground layer of the atmosphere. *Trans. Geophys. Inst. Akad. Nauk USSR*, **151**, 163–187.
- Nakanishi, M., 2000: Large-eddy simulation of radiation fog. *Bound.-Layer Meteor.*, **94**, 461–493, <https://doi.org/10.1023/A:1002490423389>.
- , and H. Niino, 2006: An improved Mellor–Yamada level-3 model: Its numerical stability and application to a regional prediction of advection fog. *Bound.-Layer Meteor.*, **119**, 397–407, <https://doi.org/10.1007/s10546-005-9030-8>.
- Overland, J. E., and Coauthors, 2016: Nonlinear response of mid-latitude weather to the changing Arctic. *Nat. Climate Change*, **6**, 992–999, <https://doi.org/10.1038/nclimate3121>.
- Pepin, N., M. Schaefer, and L. D. Riddey, 2009: Quantification of the cold air pool in Kevo Valley, Finnish Lapland. *Weather*, **64**, 60–67, <https://doi.org/10.1002/wea.260>.
- Pleim, J. E., 2007: A combined local and nonlocal closure model for the atmospheric boundary layer. Part I: Model description and testing. *J. Appl. Meteor. Climatol.*, **46**, 1383–1395, <https://doi.org/10.1175/JAM2539.1>.
- Prein, A. F., and Coauthors, 2015: A review on regional convection-permitting climate modelling: Demonstrations, prospects and challenges. *Rev. Geophys.*, **53**, 323–361, <https://doi.org/10.1002/2014RG000475>.
- Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, <https://doi.org/10.1175/MWR2906.1>.
- Rue, H., S. Martino, and N. Chopin, 2009: Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. *J. Roy. Stat. Soc.*, **71B**, 319–392, <https://doi.org/10.1111/j.1467-9868.2008.00700.x>.
- Saha, S., and Coauthors, 2010: The NCEP Climate Forecast System Reanalysis. *Bull. Amer. Meteor. Soc.*, **91**, 1015–1057, <https://doi.org/10.1175/2010BAMS3001.1>.
- Salazar, E., D. Hammerling, X. Wang, B. Sansó, A. O. Finley, and L. O. Means, 2016: Observation-based blended projections from ensembles of regional climate models. *Climatic Change*, **138**, 55–69, <https://doi.org/10.1007/s10584-016-1722-1>.
- Shin, H. H., and S.-Y. Hong, 2015: Representation of the subgrid-scale turbulent transport in convective boundary layers at gray-zone resolutions. *Mon. Wea. Rev.*, **143**, 250–271, <https://doi.org/10.1175/MWR-D-14-00116.1>.
- Sideratos, G., and N. D. Hatzigeorgiou, 2007: An advanced statistical method for wind power forecasting. *IEEE Trans. Power Syst.*, **22**, 258–265, <https://doi.org/10.1109/TPWRS.2006.889078>.
- Simpson, D. P., H. Rue, T. G. Martins, A. Riebler, and S. H. Sørbye, 2015: Penalising model component complexity: A principled, practical approach to constructing priors. *arXiv.org*, <https://arxiv.org/abs/1403.4630>.
- Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., <https://doi.org/10.5065/D68S4MVH>.
- Slingo, J. M., and T. Palmer, 2011: Uncertainty in weather and climate prediction. *Philos. Trans. Roy. Soc.*, **369A**, 4751–4767, <https://doi.org/10.1098/rsta.2011.0161>.
- Smedman, A.-S., H. Bergström, and U. Högström, 1995: Spectra, variances and length scales in a marine stable boundary layer dominated by a low level jet. *Bound.-Layer Meteor.*, **76**, 211–232, <https://doi.org/10.1007/BF00709352>.
- Steele, C. J., S. R. Dorling, R. von Glasow, and J. Bacon, 2013: Idealized WRF Model sensitivity simulations of sea breeze types and their effects on offshore wind fields. *Atmos. Chem. Phys.*, **13**, 443–461, <https://doi.org/10.5194/acp-13-443-2013>.
- , —, —, and —, 2015: Modelling sea-breeze climatologies and interactions on coasts in the southern North Sea: Implications for offshore wind energy. *Quart. J. Roy. Meteor. Soc.*, **141**, 1821–1835, <https://doi.org/10.1002/qj.2484>.
- Steenveld, G.-J., 2014: Current challenges in understanding and forecasting stable boundary layers over land and ice. *Front. Environ. Sci.*, **2**, 41, <https://doi.org/10.3389/fenvs.2014.00041>.
- Stein, M. L., 1999: *Interpolation of Spatial Data: Some Theory for Kriging*. Springer-Verlag, 228 pp.
- Steinsland, I., and H. Jensen, 2010: Utilizing Gaussian Markov random field properties of Bayesian animal models. *Biometrics*, **66**, 763–771, <https://doi.org/10.1111/j.1541-0420.2009.01336.x>.
- Sukoriansky, S., B. Galperin, and V. Perov, 2005: Application of a new spectral model of stratified turbulence to the atmospheric boundary layer over sea ice. *Bound.-Layer Meteor.*, **117**, 231–257, <https://doi.org/10.1007/s10546-004-6848-4>.
- Tao, W.-K., J. Simpson, and M. McCumber, 1989: An ice-water saturation adjustment. *Mon. Wea. Rev.*, **117**, 231–235, [https://doi.org/10.1175/1520-0493\(1989\)117<0231:AIWSA>2.0.CO;2](https://doi.org/10.1175/1520-0493(1989)117<0231:AIWSA>2.0.CO;2).
- Tegen, I., P. Hollrig, M. Chin, I. Fung, D. Jacob, and J. Penner, 1997: Contribution of different aerosol species to the global aerosol extinction optical thickness: Estimates from model results. *J. Geophys. Res.*, **102**, 23 895–23 915, <https://doi.org/10.1029/97JD01864>.
- Torn, R. D., and G. J. Hakim, 2009: Initial condition sensitivity of western Pacific extratropical transitions determined using ensemble-based sensitivity analysis. *Mon. Wea. Rev.*, **137**, 3388–3406, <https://doi.org/10.1175/2009MWR2879.1>.

- Toth, E., A. Brath, and A. Montanari, 2000: Comparison of short-term rainfall prediction models for real-time flood forecasting. *J. Hydrol.*, **239**, 134–147, [https://doi.org/10.1016/S0022-1694\(00\)00344-9](https://doi.org/10.1016/S0022-1694(00)00344-9).
- Vislocky, R. L., and J. M. Fritsch, 1995: Improved model output statistics forecast through model consensus. *Bull. Amer. Meteor. Soc.*, **76**, 1157–1164, [https://doi.org/10.1175/1520-0477\(1995\)076<1157:IMOSFT>2.0.CO;2](https://doi.org/10.1175/1520-0477(1995)076<1157:IMOSFT>2.0.CO;2).
- , and —, 1997: Performance of an advanced MOS system in the 1996–97 National Collegiate Weather Forecasting Contest. *Bull. Amer. Meteor. Soc.*, **78**, 2851–2857, [https://doi.org/10.1175/1520-0477\(1997\)078<2851:POAAMS>2.0.CO;2](https://doi.org/10.1175/1520-0477(1997)078<2851:POAAMS>2.0.CO;2).
- Wang, M., Y. Kamarianakis, A. Mahalov, M. Wagner, M. Georgescu, G. Miguez-Macho, and M. Moustouli, 2016: Spatio-temporal modeling for regional climate model comparison: Application on perennial bioenergy crop impacts. *Joint Statistical Meeting 2016*, Chicago, Illinois, Section on Statistics and the Environment, 2886–2898.
- Zeng, X., and A. Beljaars, 2005: A prognostic scheme of sea surface skin temperature for modeling and data assimilation. *Geophys. Res. Lett.*, **32**, L14605, <https://doi.org/10.1029/2005GL023030>.
- Zhang, H., Z. Pu, and X. Zhang, 2013: Examination of errors in near-surface temperature and wind from WRF numerical simulations in regions of complex terrain. *Wea. Forecasting*, **28**, 893–914, <https://doi.org/10.1175/WAF-D-12-00109.1>.