JOBNAME: MWR 00#00 2020 PAGE: 1 SESS: 8 OUTPUT: Thu Mar 26 16:56:31 2020 Total No. of Pages: 19
/ams/mwr/0/mwrD190321

Sensitivity Analysis of the Spatial Structure of Forecasts in Mesoscale Models: Noncontinuous Model Parameters

CAREN MARZBAN

Applied Physics Laboratory, and Department of Statistics, University of Washington, Seattle, Washington

ROBERT TARDIF

Department of Atmospheric Sciences, University of Washington, Seattle, Washington

SCOTT SANDGATHE

Applied Physics Laboratory, University of Washington, Seattle, Washington

(Manuscript received 30 September 2019, in final form 5 February 2020)

ABSTRACT

In a recent work, a sensitivity analysis methodology was described that allows for a visual display of forecast sensitivity, with respect to model parameters, across a gridded forecast field. In that approach, sensitivity was assessed with respect to model parameters that are continuous in nature. Here, the analogous methodology is developed for situations involving noncontinuous (discrete or categorical) model parameters. The method is variance based, and the variances are estimated via a random-effects model based on 2^{k-p} fractional factorial designs and Graeco-Latin square designs. The development is guided by its application to model parameters in the stochastic kinetic energy backscatter scheme (SKEBS), which control perturbations at unresolved, subgrid scales. In addition to the SKEBS parameters, the effect of daily variability and replication (both, discrete factors) are also examined. The forecasts examined are for precipitation, temperature, and wind speed. In this particular application, it is found that the model parameters have a much weaker effect on the forecasts as compared to the effect of daily variability and replication, and that sensitivities, weak or strong, often have a distinctive spatial structure that reflects underlying topography and/or weather patterns. These findings caution against fine-tuning methods that disregard 1) sources of variability other than those due to model parameters, and 2) spatial structure in the forecasts.

1. Introduction

Sensitivity analysis (SA) generally refers to methods for assessing how inputs of a process affect its output, with the terms "inputs," "process," and "output" interpreted in an abstract sense. For example, Lucas et al. (2013) consider climate models and perform SA to explore the effect of model parameters on the probability of model crashes. The well-known adjoint method (Errico 1997) can also be viewed as a SA method, although the main goal there is fine-tuning or calibration of model parameters (Ancell and Hakim 2007; Safta et al. 2015; Hacker et al. 2011; Laine et al. 2012; Ollinaho et al. 2014). Sometimes SA is performed not necessarily for the purpose of calibration but to simply shed light on the underlying physical processes (Roebber 1989; Roebber and Bosart 1998; Robock et al. 2003; Marzban 2013; Marzban et al. 2014).

The subtle but important differences between SA, fine-tuning, and calibration are discussed by Marzban et al. (2018a), where an (object-oriented) SA method is developed to determine how model parameters affect various features of spatially coherent "objects" in forecast fields. The SA method developed here belongs to that latter class; that is, wherein the main aim is not fine-tuning or data assimilation [as in Ancell and Hakim (2007), Järvinen et al. (2012), and Laine et al. (2012)], but rather to obtain some sense of sensitivity for the purpose of understanding how model parameters affect the forecasts (Aires et al. 2014; Fasso 2006; Marzban et al. 2019, 2018a, 2014; Oakley and O'Hagan 2004; Saltelli et al. 2010; Sobol' 1993; Zhao and Tiede 2011). Of course, one may use that understanding to improve forecasts, but that is only a secondary goal.

Corresponding author: Caren Marzban, marzban@stat.washington.edu

DOI: 10.1175/MWR-D-19-0321.1

JOBNAME: MWR 00#00 2020 PAGE: 2 SESS: 8 OUTPUT: Thu Mar 26 16:56:31 2020 Total No. of Pages: 19
/ams/mwr/0/mwrD190321

This paper is a member of a sequence of papers all dealing with SA of gridded forecasts in numerical weather prediction models. In the first of the sequel (Marzban et al. 2018a), an object-oriented SA method was developed to assess the effect of 11 models parameters in COAMPS (Hodur 1997) on the characteristics of spatially coherent ''objects'' in the forecast. In a follow-up paper (Marzban et al. 2018b), the effect of the same model parameters on the spatial structure of forecasts was examined. Common to both of those works was the sampling scheme used for selecting the values of the model parameters; given that the model parameters were all continuous quantities, Latin Hypercube Sampling (LHS) was employed (further described in the method section).

In extending the developed methodology to incorporate noncontinuous (e.g., discrete or categorical) model parameters, Marzban et al. (2019) examine the effect of two model parameters in stochastic kinetic energy backscatter schemes (SKEBS) (Leith 1990; Mason and Thomson 1992; Berner et al. 2008, 2011; Shutts 2005) on forecasts made by the Weather Research and Forecasting (WRF) Model (Skamarock and Klemp 2008). In that specific application, although the two model parameters (amplitude of perturbations to potential temperature and nondivergent wind) are continuous, the SA included two additional factors that are discrete—they measure how the sensitivities vary across 1) days, and 2) replication of SKEBS. The existence of discrete factors in the SA precludes LHS, and so the sampling of the parameter space was done according to the Graeco-Latin square design (GLSD), further explained in the method section. That analysis was object oriented because the main focus was on the characteristics of spatially coherent structures in the forecasts. The object features examined included the number, size, and intensity of upper-air jet streaks, low-level jets, precipitation areas, and frontal boundaries (i.e., baroclinic zones). A complex pattern of effects was found, but an unambiguous conclusion was that the object features are much less sensitive to perturbations of the model parameters in comparison to daily variability and variability due to replication.

As such, what remains to be done is the development of a methodology for assessing the impact of noncontinuous model parameters on the spatial structure of forecasts (in SKEBS/WRF). Given the focus on spatial structure, the approach taken here is similar to that of Marzban et al. (2018b) wherein ''maps'' of sensitivity are generated to visually convey the spatial pattern of sensitivity. However, the extension of the SKEBS analysis to more than two model parameters, and parameters that include a mix of continuous and noncontinuous quantities, requires an alternative design that in some ways is even more ''efficient'' than GLSD. The design is generally referred to as a fractional factorial design (FFD), described in the method section. Here, suffice it to say that FFD is more efficient than GLSD because the necessary number of runs for estimating the sensitivities is smaller in FFD (Montgomery 2009).

The main goals of this paper are 1) to develop a spatial SA method wherein the factors of interest are not continuous, 2) illustrate the relevance (to SA) of some well-established results from the field of experimental design, and 3) demonstrate the methodology through a specific example.

The outline of the paper is as follows: Given that the aforementioned SKEBS analysis utilized a GLSD, here the analysis also begins with a GLSD; but by contrast, focus is on spatial structure of the forecast field, not features of objects therein. The first forecast field analyzed is that of precipitation, again because that was the field examined in the prior work. These results are presented as a means of connecting the prior work to the main focus of the work reported here, which is to employ a FFD to assess the sensitivity of precipitation, wind speed, and temperature, with respect to daily variability, variability due to replication, and that due to eight model parameters in SKEBS/WRF. Although, forecasts at 3-h intervals (between 0 and 120 h) are examined using the developed methodology, only the results pertaining to 3- and 24-h forecasts are reported here. Although several conclusions are drawn regarding specific model parameters and forecast fields, the broad conclusions of earlier studies (e.g., COAMPS with 11 model parameters, or SKEBS with two model parameters) are found here again, that is, 1) the model parameters have a very small effect relative to effects of daily variability and replication, and 2) even with small effects, many of the model parameters have a distinct spatial structure of sensitivities. Most of the observed spatial structures are not readily explainable (e.g., in terms of topography). It is hypothesized that the spatial structure observed in the sensitivities is a consequence of that in the underlying topography and/or weather patterns on the specific days included in the data. All of these findings and conjectures are consistent with (and confirm) claims that fine-tuning model parameters must account for variability across space, weather systems, and other sources of variability.

2. Method

All of the previously mentioned SA methods are variance based, which simply means that the effect of a factor (e.g., model parameter) on the response

JOBNAME: MWR 00#00 2020 PAGE: 3 SESS: 8 OUTPUT: Thu Mar 26 16:56:31 2020 Total No. of Pages: 19
/ams/mwr/0/mwrD190321

(e.g., forecast of temperature) is measured by the percentage of the variability in the latter that can be explained by the former. As such, variance plays an important role. The methodology developed here is comprised of relatively standard techniques in the field of experimental design (Montgomery 2009). There are two main ingredients: 1) the statistical model for estimating sensitivities, and 2) the sampling scheme for generating the data necessary for the estimation.

a. Random-effects models

For the purpose of estimating sensitivities linear models are often sufficient because even if the underlying processes involve nonlinear relations, to first approximation they are linear. The model employed here is

$$y_{i,j,\ldots k,l} = \mu + \text{day}_i + X1_j + \cdots + Xk_k + \varepsilon_{i,j,\ldots k,l}, \qquad (1)$$

where the response $y_{ijk,\ldots,l}$ denotes a measurement of some quantity of interest (e.g., amount of precipitation at a grid point) on the ith day, for the jth, . . . , kth values of the model parameters $X1$, $X2$, . . . , Xk, respectively, and for the lth replication of the experiment. The term μ denotes the true mean of the response across all values of the factors, and ε accounts for any factor that has not been included in the model—often called "error." The index l varies over the number of replications (i.e., the number of times the entire experiment has been repeated). As described in the data section, the number of days in this study is nine, and the entire experiment is replicated six times.

It is worth mentioning that the model in Eq. (1) is not a regression model mapping model parameters to forecasts, because that would require the predictors (i.e., model parameters) to be continuous, or at least numerical. The "predictors" appearing on the right-hand side of the equation represent the mean of the response y at different levels of the factors. In fact, the factors in that equation are not even restricted to be numerical quantities; they may be categorical (e.g., yes/maybe/no). Models of this kind are often referred to simply as linear models (or ANOVA-type models) (Montgomery 2009).

The more common use of linear models treats the factors on the right-hand side of Eq. (1) as fixed (nonrandom) quantities, with the exception of the error term, which is assumed to be a zero-mean, normally distributed random variable with variance σ^2_ε. Such models are called fixed-effects models, and statistical tests exist for testing whether any of the factors, or a given factor, has an effect on the response y. In fixed-effects models, the results of the tests apply to only the *specific values/levels* of the factors appearing in the data. For example, if the $X1$ factor is found to have a small p value, all one can conclude is that there is evidence that the true mean of the response varies across the specific values/levels taken by that factor in the data. To generalize that conclusion to *all possible values/levels*, one must treat the factor as a random variable.

In a random-effects model, all of the terms on the right-hand side of Eq. (1) (except μ) are random variables. The simplest probability model for these random variables is that they are zero-mean, normally distributed variables, with corresponding variances satisfying

$$\sigma^2_{\text{response}} = \sigma^2_{\text{day}} + \sigma^2_{X1} + \sigma^2_{X2} + \cdots + \sigma^2_\varepsilon, \qquad (2)$$

where the so-called variance components (on the right side of the expression) are to be estimated from data. A natural quantity in random-effects models is the intraclass correlation (ρ), defined as the ratio of each variance component to the total variance $\sigma^2_{\text{response}}$; it conveys the proportion of the total variability in the response that can be explained by each factor in the model. This intraclass correlation is the measure of sensitivity used in the present work.

Two common estimators for the variance components are the analysis-of-variance, and the restricted-maximum-likelihood estimators (Montgomery 2009); the former is the simpler of the two, and is accompanied by analytic formulas for computing confidence intervals, but it has the defect of sometimes leading to negative values for the estimates of the variance components. For this reason, the latter estimator is used. The confidence intervals for ρ are computed via a randomization procedure whereby the observed forecast value at a given grid point is randomly assigned to a different/random setting of the factors; the collection of these ρ values across all the grid points constitutes the empirical sampling distribution of ρ (from which confidence intervals and p values can be found).

In the present application, where spatial maps of sensitivity are produced, it is difficult (or even impossible) to produce a map that simultaneously displays sensitivity *and* its confidence interval. Instead, only the 95% lower confidence bound (LCB) for ρ is shown. A focus on the LCB is appropriate because the question of interest is How *small* can ρ be? Asked differently, What is the *smallest* plausible ρ consistent with the observed data? Switching from a point estimate to an interval estimate renders the analysis statistically more rigorous; for example, one can say with 95% confidence, that the true value of ρ at a given grid point is larger than that observed. However, for the purpose of establishing the existence of spatial structure, one may use either estimate, because the relationship between the point

JOBNAME: MWR 00#00 2020 PAGE: 4 SESS: 8 OUTPUT: Thu Mar 26 16:56:31 2020 Total No. of Pages: 19
/ams/mwr/0/mwrD190321

estimate and the LCB is monotonic. Consequently, the existence of spatial structure in a map of one estimate implies a spatial structure in the map of the other estimate. For this reason, both estimates are used here.

It is worth mentioning that the LCB for ρ cannot be used for testing the hypothesis that $\rho = 0$, because $0 < \text{LCB} < 1$ (strictly), by construction. For the same reason, the upper confidence bound for ρ cannot be used for testing the hypothesis $\rho = 1$. Given that the LCB is not used for hypothesis testing, corrections for multiple hypothesis testing (at multiple grid points) are not necessary; it is the *spatial structure* of the LCB that is of interest. This technical point is further addressed in the discussion section.

The formula in Eq. (2) is standard in statistical texts on experimental design. There, σ_ε^2 represents the variance of the response with respect to any factors not included in the model—hence the name "error variance." However, in data generated from a computer experiment—often called "computer data" (Bowman et al. 1993; Sacks et al. 1989; Welch et al. 1992)—the only sources of variability are those implicit in the design of the experiment. For the design in the present study, the only sources of variability are from changing the model parameters, different days, and different replications. In other words, for any specific values of the eight model parameters, on any specific day, and for a specific replication, running SKEBS/WRF will lead to the same response. That property is a defining characteristic of computer data. Therefore, for the model in Eq. (1), one has $\sigma_\varepsilon^2 = \sigma_{\text{rep}}^2$ (i.e., variance due to replication can be estimated from the variance of the errors).

A possible violation of this equality would occur if/when there are interactions between the factors. The introduction of interaction terms in Eq. (1) leads to an underdetermined model, and so the effects cannot be uniquely estimated (Montgomery 2009), a shortcoming of both GLSD and FFD. In that situation, main effects and interaction effects become *aliased,* and only their sum can be estimated from data. Therefore, assuming that the interaction effects are relatively small, the interpretation of the error variance as variance due to replication is permissible. Justifications for this assumption are discussed in section 2b.

b. Sampling design

As mentioned in the introduction, for continuous model parameters the sampling method of choice is often the Latin hypercube sampling (LHS) (Cioppa and Lucas 2007; Marzban 2013); it has been used in sampling the model parameter space (Marzban et al. 2014), selecting the members in ensembles for ensemble forecasting (Hacker et al. 2011), for emulation (Santner et al. 2003), and for performing variance-based SA (Saltelli et al. 2010, 2008). The popularity of this sampling scheme derives from the fact that it leads to estimates that are often more precise (never, less precise) than simple random sampling (Cioppa and Lucas 2007).

When the model parameters are discrete or categorical, LHS does not apply, and the choice of the sampling scheme is more complicated because there exists a wide range of schemes optimized for different circumstances. Many sampling schemes are designed to minimize the number of runs necessary for estimating the parameters of interest–an important consideration given the computationally expensive nature of numerical models. Two of the most common such sampling schemes are called Graeco-Latin square designs (GLSD), and 2^{k-p} fractional factorial designs (FFD). Details of GLSD can be found in many texts on experimental design (e.g., Montgomery 2009), but they are also described briefly in the appendix. The GLSD is also the method that was used in Marzban et al. (2019). The attractiveness of the GLSD originates from the fact that the necessary number of runs is significantly smaller than in a full factorial design. For example, given k factors (e.g., model parameters), each taking L possible values, a full factorial design would require L^k runs; by contrast GLSD requires L^2 runs, regardless of the number of factors. Consequently, GLSD is desirable for handling large number of factors.

Here, in addition to GLSD, FFD is also used–in fact, used more–because it is even more efficient. In FFD each factor takes only two values, often set to the minimum and maximum possible values of the factor. As such, the necessary number of runs is 2^k. Although the binary treatment of the factors does considerably reduce the necessary number of runs, the "magic" of the FFD is in its ability to estimate the effect of the factors with a fraction of the 2^k runs. The fractions are often 1/2, 1/4, 1/8, etc., leading to these designs being called 2^{k-p} designs.

Not all values of k and p, however, are desirable because they often lead to aliasing of effects wherein the effect of one factor cannot be disentangled from the effect of another factor. However, special values of k and p have been discovered that do allow for the estimation of the main effects; in these special solutions, even when aliasing does occur, it involves high-order interactions which are generally much smaller than main effects. That interaction effects are generally much smaller than main effects is borne out due to several "principles:" the principle of hierarchical ordering, the principle of effect sparsity, and the principle of effect hierarchy [see pp. 192, 230, 272, 314, 329 in Montgomery (2009), and 33–34 in Li et al. (2006)].

JOBNAME: MWR 00#00 2020 PAGE: 5 SESS: 8 OUTPUT: Thu Mar 26 16:56:32 2020 Total No. of Pages: 19
/ams/mwr/0/mwrD190321

These special solutions are generally listed in texts on experimental design, or encoded in computer software. For example, appendix X in Montgomery (2009) reveals T1 the special values of k and p in Table 1 here. The first column of Table 1 implies that in a problem involving $k = 8$ binary factors, one can estimate the main effects with only $2^{8-4} = 16$ runs—a sharp reduction from $2^8 = 256$ runs of the model in a full factorial design. Even more dramatic reduction in the number of runs is obtained if the problem involves more factors. The last column in Table 1 implies that in a problem involving 15 binary factors, the main effects can be estimated with only $2^{15-11} = 16$ runs, as compared to the $2^{15} = 32\,768$ runs one may (be tempted to) perform in a full factorial design.

These special designs require running a specific subset of the 2^k possible runs, and these subsets too are often found in texts and computer software on experimental design. For the present study involving eight model parameters, the 2^{8-4} design requires the fewest number of runs, and the special values of the eight model parameters are shown in T2 Table 2.

For the present application, the eight SKEBS/WRF T3 parameters are shown in Table 3, and they are furthered discussed in the data section. It is the discrete nature of the last four parameters that calls for the GLSD and/or FFD as opposed to designs based on LHS.

c. *Organization of the experiments*

To maintain continuity with previous work, first a GLSD experiment is performed on the four factors used in the object-oriented SA work in (Marzban et al. 2019); they are Day, Replication, and Par1 and Par2 in Table 3. It may be worth repeating that the difference between the current and the past SKEBS work is that the former examines the sensitivity maps across the entire grid field, while the latter examined the sensitivity of objects within the forecast field. After the GLSD experiment, all eight model parameters are examined in a 2^{8-4} FFD. The Day factor is still included in the model, and so there are a total of nine factors in the model of Eq. (1).

3. Data/application

To generate the computer data necessary for performing the SA, version 3.7.0 of the WRF-ARW Model with lateral boundary conditions specified every 6 h from output of the Global Forecast System (GFS) is used. All of the standard WRF parameters are the default "out of box" parameters, with a 25-km domain over the continental United States. Nine days are selected between December 2014 and February 2015, with initial forecast hours 10 days apart to ensure minimal temporal association between days. Specifically, they are 1, 11, 21, 31 December 2014; 10, 20, 30 January; and 9, 19 February 2015. Winter months were chosen for the high degree of variability with regards to the forecast fields examined here (i.e., precipitation, temperature, and wind speed). Forecasts of 3-h-accumulated precipitation were generated in 3-h intervals, although only the 3- and 24-h forecasts are analyzed here. Similarly, temperature forecasts at only 3 and 24 h are considered; although both 2-m and 500-hPa forecasts were analyzed, only the former are presented here. As for wind speed, 3- and 24-h forecasts at 250 hPa are studied here, although results at 850 hPa are also available.

TABLE 1. Special values of k and p that allow estimation of main effects (i.e., the effect of model parameters on the response) in a total of 2^{k-p} runs.

k	8	8	8	10	10	11	12	13	14	15
p	4	3	2	6	5	7	8	9	10	11
2^{k-p}	16	32	64	16	32	16	16	16	16	16

In the prior GLSD experiment involving four factors, because one of the factors was Day, and because it took nine values, Par1 and Par2 were discretized into nine values as well (GLSD requires the same number of values/levels for all factors). For Par1 they are (0.5×10^{-5}, 1.6875×10^{-5}, 2.875×10^{-5}, 4.0625×10^{-5}, 5.25×10^{-5}, 6.4375×10^{-5}, 7.625×10^{-5}, 8.8125×10^{-5}, 1.0×10^{-4}), and for Par2 they are (0.5×10^{-5}, 1.6875×10^{-5}, 2.875×10^{-5}, 4.0625×10^{-5}, 5.25×10^{-5}, 6.4375×10^{-5}, 7.625×10^{-5}, 8.8125×10^{-5}, 1.0×10^{-4}). The range of the nine values are chosen to contain the recommended SKEBS values, but span one order of magnitude smaller and one order of magnitude larger than the default values. It is worth repeating that in random-effects models, inference of the sensitivities pertains to *all* possible values of the parameters, not just to the specific values selected in the data; for this reason, the specific nine values selected here do not play an important role in the final analysis.

In the 2^{k-p} FFD experiment, all eight model parameters are treated as binary, taking the minimum and maximum values shown in Table 3.

One of the main goals here is to assess the effect of the replication factor (i.e., the purely stochastic component of SKEBS/WRF) and how it compares with the effect of the other factors (i.e., the Day factor and the eight model parameters). Here, the GLSD and the FFD experiments are each replicated six times.[1] To clarify the two designs,

[1] In the jargon of experimental design, one says that the GLSD experiment is a three-factor design (i.e., Day, Par1, Par2) crossed with the replication factor.

JOBNAME: MWR 00#00 2020 PAGE: 6 SESS: 8 OUTPUT: Thu Mar 26 16:56:32 2020 Total No. of Pages: 19
/ams/mwr/0/mwrD190321

TABLE 2. The 16 runs corresponding to the 2^{8-4} design, involving eight factors. These runs assure that the effect of the model parameters can be estimated independently of the existence of any two-way interactions.

*X*1	−1	1	−1	1	−1	1	−1	1	−1	1	−1	1	−1	1	−1	1
*X*2	−1	−1	1	1	−1	−1	1	1	−1	−1	1	1	−1	−1	1	1
*X*3	−1	−1	−1	−1	1	1	1	1	−1	−1	−1	−1	1	1	1	1
*X*4	−1	−1	−1	−1	−1	−1	−1	−1	1	1	1	1	1	1	1	1
*X*5	−1	−1	1	1	1	1	−1	−1	1	1	−1	−1	−1	−1	1	1
*X*6	−1	1	−1	1	1	−1	1	−1	1	−1	1	−1	−1	1	−1	1
*X*7	−1	1	1	−1	1	−1	−1	1	−1	1	1	−1	1	−1	−1	1
*X*8	−1	1	1	−1	−1	1	1	−1	1	−1	−1	1	1	−1	−1	1

consider the number of runs: At every forecast hour, the total number of model runs is $6 \times 9^2 = 486$ for GLSD; that is, six replications of 9^2 runs of a GLSD involving the factors Day, Par1, and Par2. For the 2^{8-4} FFD the total number of model runs is $6 \times 9 \times 16 = 864$ (i.e., 6 replications, 9 days, and 16 runs involving the 8 model parameters). Although these numbers may seem large, they are significantly smaller than what would be necessary in a full factorial design: $6 \times 9^3 = 4374$, and $6 \times 9 \times 2^8 = 13\,824$, respectively.

4. Results

In a variance-based approach to SA, one varies all of the factors of interest (Day, Replication, and model parameters) according to some sampling scheme (e.g., GLSD or FFD), runs the forecasting model (here WRF/SKEBS) with those settings, computes the resulting variance for the response variable (e.g., accumulated precipitation at a given grid point), and then uses statistical models of the type in Eq. (1) to apportion that variability across the factors in the model, via Eq. (2).

To get a sense of the magnitude and nature of these sources of variability, Fig. 1 shows 24-h forecasts of accumulated precipitation (in mm), for the 9 days examined here, with the model parameters set to their default values. Figure 2 shows the 24-h forecasts of accumulated precipitation for one perturbation (×3 in Table 2) of the eight model parameters according to FFD. Evidently, across the perturbations, the spatial structure of the forecasts changes in a complex fashion. The 16 perturbations in Table 2 lead to even more varied and complex changes, depending on day, and the effect of replication (not shown). The main purpose of the methodology developed here is to determine the extent to which each model parameter contributes to these changes.

To make connection with prior work, the first SA result shown here is from a GLSD involving only the four factors examined in Marzban et al. (2019), and their effect on 3-h forecasts of accumulated precipitation. The top-left panel in Fig. 3a shows the total variance (across days, parameter values, and replication) [i.e., $\sigma^2_{\text{Response}}$ in Eq. (2)], estimated at each grid point. Darker (lighter) shades of gray correspond to more (less) variability. It is evident that there exists significant spatial structure. Some of the causes for the spatial structure are discussed below, but the main question here is How much of that spatial structure is due the various factors that altogether account for the observed variability? The remaining panels in Fig. 3a show maps of intraclass correlation ρ, gauging the proportion of total variability that can be attributed to each of the factors; a high/low (dark/light) value implies high/low sensitivity. The numbers appearing atop each panel are the spatial average of the ρ values across the panel. That the Day panel is mostly black, and that a relatively large proportion (about 75%) of the total variability can be attributed to daily variability, is a direct consequence of the significant daily

TABLE 3. The eight model parameters (Par) (and their range) whose sensitivity is assessed.

	SKEBS name	Description	Range
Par1	tot backscat t	Total backscattered dissipation rate for potential temperature	$(5 \times 10^{-7}, 1 \times 10^{-4})$
Par2	tot backscat psi	Total backscattered dissipation rate for streamfunction	$(5 \times 10^{-7}, 1 \times 10^{-4})$
Par3	rexponent t	Spectral slope of potential temperature perturbations	(−1.83, −0.01)
Par4	rexponent psi	Spectral slope of streamfunction perturbations	(−1.83, −0.01)
Par5	ztau t	Decorrelation time (s) for potential temperature perturbations	(10 800, 43 200)
Par6	ztau psi	Decorrelation time (s) for streamfunction perturbations	(10 800, 43 200)
Par7	lmaxforct	Maximal forcing wavenumber in latitude for potential temperature perturbations	(1, 65)
Par8	stoch vertstruc opt	Constant vertical structure of random pattern generator	(0 or 1)

JOBNAME: MWR 00#00 2020 PAGE: 7 SESS: 8 OUTPUT: Thu Mar 26 16:56:32 2020 Total No. of Pages: 19
/ams/mwr/0/mwrD190321

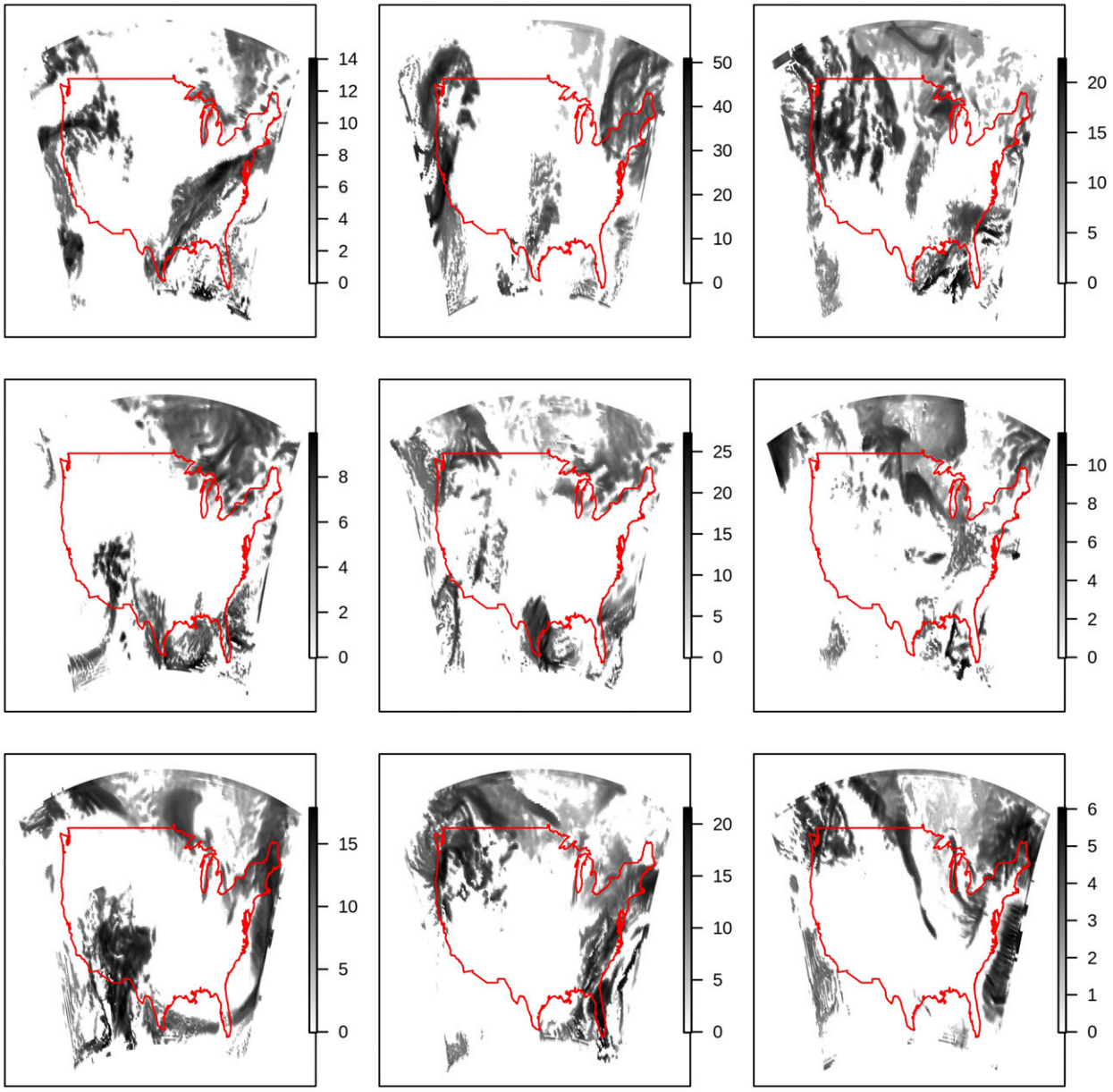

FIG. 1. The 24-h forecasts of 3-h-accumulated precipitation (in mm) for the 9 days examined here, with the SKEBS model parameters set to their default values.

variability seen in Fig. 1. The white patches correspond to grid points where no precipitation was forecast on any of the days, for any of the model parameter values, and for any replication.

The finding that a relatively large proportion (25%) of total variability is due to replication, and that a very small proportion (0.1%) is due to the two model parameters, is consistent with the conclusions of Marzban et al. (2019) where features of forecast objects were the response variable of the SA. Less expected is the nontrivial spatial structure associated with Replication and Par2. It may be tempting to attribute the dark region in the corresponding panels (in the northeast of the domain) to the persistence of precipitation across the 9 days in the data; however, an examination of Fig. 1 reveals that, in fact, precipitation is only mildly persistent in that region. But regardless of its explanation, one important consequence of this spatial structure is that the replication of SKEBS/WRF ought not be expected to have a uniform effect across the spatial domain.

Figure 3b shows the analogous results for 24-h forecasts of accumulated precipitation. The most notable

JOBNAME: MWR 00#00 2020 PAGE: 8 SESS: 8 OUTPUT: Thu Mar 26 16:56:33 2020 Total No. of Pages: 19
/ams/mwr/0/mwrD190321

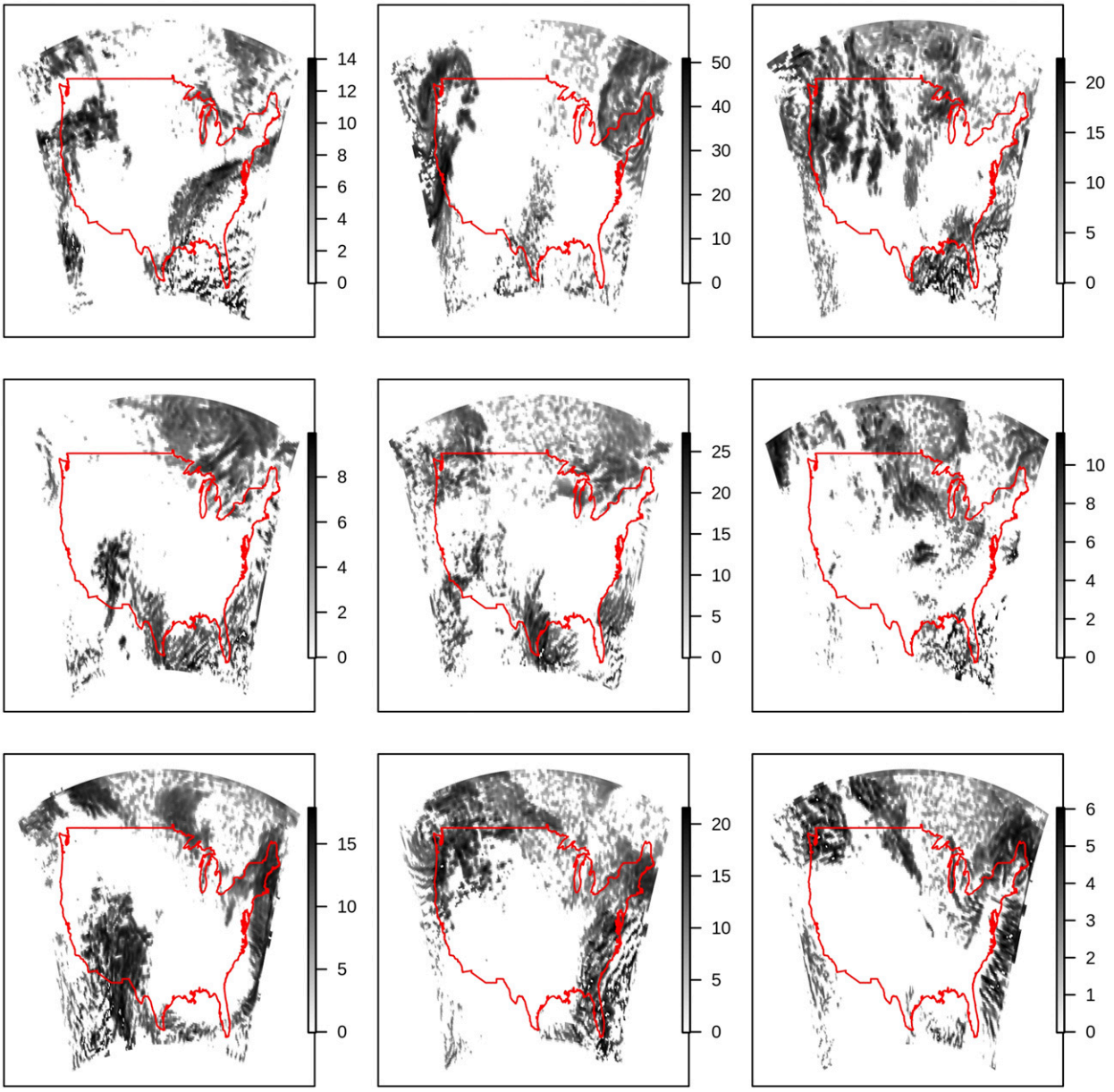

FIG. 2. As in Fig. 1, but with the eight SKEBS model parameters perturbed according to FFD (×3 in Table 2).

difference is that the solid/white patches in Fig. 3a are not present in Fig. 3b; this is because at 24 h, every grid point does receive some level of precipitation, and so, it is possible to assess sensitivity at all grid points. Another noteworthy difference is that whereas the overall contribution of Day and Replication factors, at 3 h, are about 75% and 25%, respectively, at 24 h, the contributions are reversed—32% and 68%, respectively. The increasing contribution—from 25% to 68%—of the Replication factor over Day (weather variability across our sample) is a consequence of the nonlinear growth of the stochastic perturbations related to flow instabilities as model forecasts are integrated forward. The decrease from 75% to 32% is then a simple consequence of the fact that the sum of the contributions across all factors must be 100%.

Perhaps even more importantly, the coherent spatial structures seen in all four panels are far more diffused at 24 h compared to the 3-h results. The reason for this phenomenon is probably that the model is in its spinup phase early during this period, and that a longer forecast horizon allows for a greater influence of SKEBS perturbations on simulated precipitation. Also, the less diffused spatial structure at 3 h is a consequence of the

JOBNAME: MWR 00#00 2020 PAGE: 9 SESS: 8 OUTPUT: Thu Mar 26 16:56:33 2020 Total No. of Pages: 19
/ams/mwr/0/mwrD190321

(coincidental) similarities between the actual forecast fields across the 9 days (Fig. 1); at 24 h the different movement of the weather systems tends to decorrelate the sensitivity fields, leading to the more diffused appearance of the sensitivity maps. In summary, at 24 h, a very small fraction (about 0.2%) of the variability in precipitation forecasts is due to model parameters, and there exists a relatively diffused spatial structure in the sensitivities.

A comment about spatial structure is in order. Spatial structure is inherently a multifaceted concept, and consequently difficult to quantify. Throughout this paper, the spatial structure is assessed only qualitatively, and through the visual examination of sensitivity maps. Attention is placed on the degree to which the image is ''clumpy'' versus ''diffused.'' This is not a limitation in the present work, because the spatial structure is invariably addressed in the context of comparing one field with another. For example, it is quite evident that in Fig. 3b the panels corresponding to Day and Replication factors have similar (even complementary) spatial structure, and that both of those structures are distinct from that seen in the panels for Par1 and Par2, which themselves have similar spatial structure. Although a proposal to quantify these spatial structures is presented in the discussion section, the remainder of this paper will continue to address spatial structure in a qualitative and comparative sense.

Also, it is important to emphasize that in spite of the resemblance between many of the aforementioned spatial patterns (e.g., Fig. 3b) and weather patterns (e.g., Fig. 1) the former are patterns in the *sensitivity* of the forecasts. Specifically, if a region in Fig. 3b is mostly dark colored, then it follows that the forecast in that region is highly sensitive to changes in the corresponding model parameter. It does not follow that increasing (decreasing) the model parameter will lead to higher (lower) forecast values. As such, the spatial patterns in these figures are difficult to explain in terms of meteorological processes because complex nonlinear interactions determine the growth (or decay) rates of introduced perturbations. As seen below (for temperature forecasts), some spatial patterns do lend themselves to simple explanations (following topography, for example), but in general, or at least in the case of precipitation, simple explanations of the patterns are not readily available. Therefore, throughout this paper, focus is placed on establishing the existence of spatial structure in the map of sensitivities, and less effort is made to explain the patterns in meteorological terms.

F4 Returning to the analysis, Figs. 4a and 4b are the analogs of Figs. 3a and 3b but from an FFD involving all eight model parameters in Table 3. The top-left panel shows the variability in accumulated precipitation across the days, replications, and perturbations of the model parameters. Again, the main goal of the SA method developed here is to ''decompose'' this figure into its variance components. However, whereas in Fig. 3 the quantity being displayed is the point estimate of ρ, the quantity displayed in Fig. 4 is the 95% LCB of ρ. As discussed in the method section, in spite of the statistical rigor accompanying the LCB, examination of the point estimate of ρ itself can also be revealing. Based on Fig. 4a, it appears that the aforementioned spatial pattern associated with Par2 (see Fig. 3a) is present for Par3 and Par7 to an even stronger extent. An examination of the point estimates of ρ itself (not shown) reveals that Par5 has a similar spatial structure. By contrast, Par1, Par4, Par6, and Par8 are associated with a far more diffused spatial structure.

Note that in Fig. 4a the spatial patterns associated with the total variability, Replication, Par1 and Par2 are nearly identical to those found in the GLSD experiment (Fig. 3). Such agreements are not guaranteed because there are several important differences between the two designs. First, the model parameter perturbations in the GLSD are different from those in the FFD. Second, whereas the Day factor is one of the factors in the GLSD, in FFD it is not. Consequently, the similarity of the spatial structures across the two designs is a testament to the robustness of the results. However, these differences in design do lead to some small-scale differences in the sensitivity maps, and so the agreement between the GLSD and FFD results is not exact.

To summarize Fig. 4a, based on the domain averages shown atop each panel, 3-h forecasts of accumulated precipitation are affected mostly by daily variability and replication (in that order), and almost not at all by any of the model parameters. Parameters 2, 3, 5, and 7, in spite of having negligible contribution to sensitivity, appear to have similar spatial structures, and that structure is distinct from that associated with the other model parameters.

For 24-h forecasts, the results are shown in Fig. 4b. The spatial structures seen at 3 h due to daily variability and replication are still clearly visible. The overall contribution of all the model parameters is only about 2% of the total variability, and the spatial structure of sensitivities for all of them is relatively diffused.

The sensitivity maps for 3-h forecasts of 2-m temperature (not shown) display a map of total variance that suggests relatively high sensitivity over land, but much less sensitivity over the oceans. The decomposition of that variability leads to a nearly homogeneous map for the daily contribution, and the replication map displays

JOBNAME: MWR 00#00 2020 PAGE: 10 SESS: 8 OUTPUT: Thu Mar 26 16:56:33 2020 Total No. of Pages: 19
/ams/mwr/0/mwrD190321

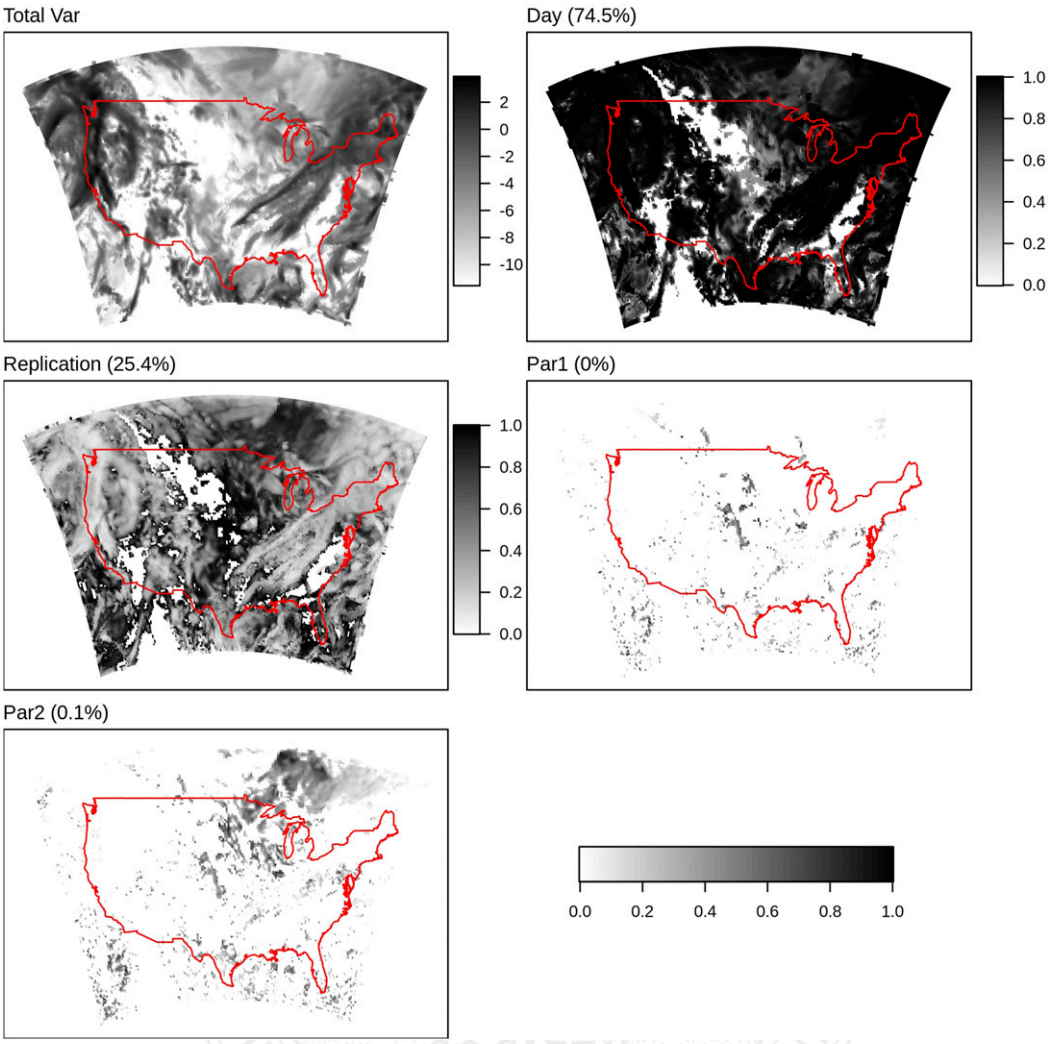

FIG. 3. (a) For 3-h forecasts of accumulated precipitation, the spatial structure of total variability (across Day, Replication, and two model parameters) according to a GLSD, and its decomposition into variance components. Darker (lighter) shades of gray correspond to higher (lower) values of the intraclass correlation ρ. The gray bar in the bottom right refers to the ρ values appearing in the panels for the model parameters. (b) As in (a), but for 24-h forecasts of accumulated precipitation.

sensitivity over the water of Pacific Ocean and the Gulf of Mexico. The maps corresponding to the model parameters show no coherent spatial structure, at all. (This is the reason why the corresponding figures for 2-m temperature are now shown.)

The analogous results for 24-h forecasts are shown in F5 Fig. 5; these are the point estimates of ρ, because the LCB of ρ leads to maps that are nearly white when viewed on a common grayscale. The total variance of this response variable has a distinct spatial structure, with larger variance over land, and much less so over the Pacific Ocean and the Gulf of Mexico. The main task of the proposed methodology is to assess how that variability is apportioned across the factors examined here. It can be seen that on the average nearly all of the variability (92%) is due to daily variability, with the Pacific Ocean and the Gulf of Mexico still showing lower sensitivity than the remainder of the domain. A much smaller proportion (7%) of total variability is due to replication, but this time the aforementioned waters show more sensitivity. Par3 and Par7 have a similar spatial structure wherein their effects appear to be restricted to the aforementioned waters. Par4 and Par6 appear to affect the northern and southern regions, respectively, of the Pacific Ocean. Par1 appears to affect the eastern coast of the United States and the waters off that coast. Again, here, less effort is made to explain these patterns; the main goal is to demonstrate the

JOBNAME: MWR 00#00 2020 PAGE: 11 SESS: 8 OUTPUT: Thu Mar 26 16:56:34 2020 Total No. of Pages: 19
/ams/mwr/0/mwrD190321

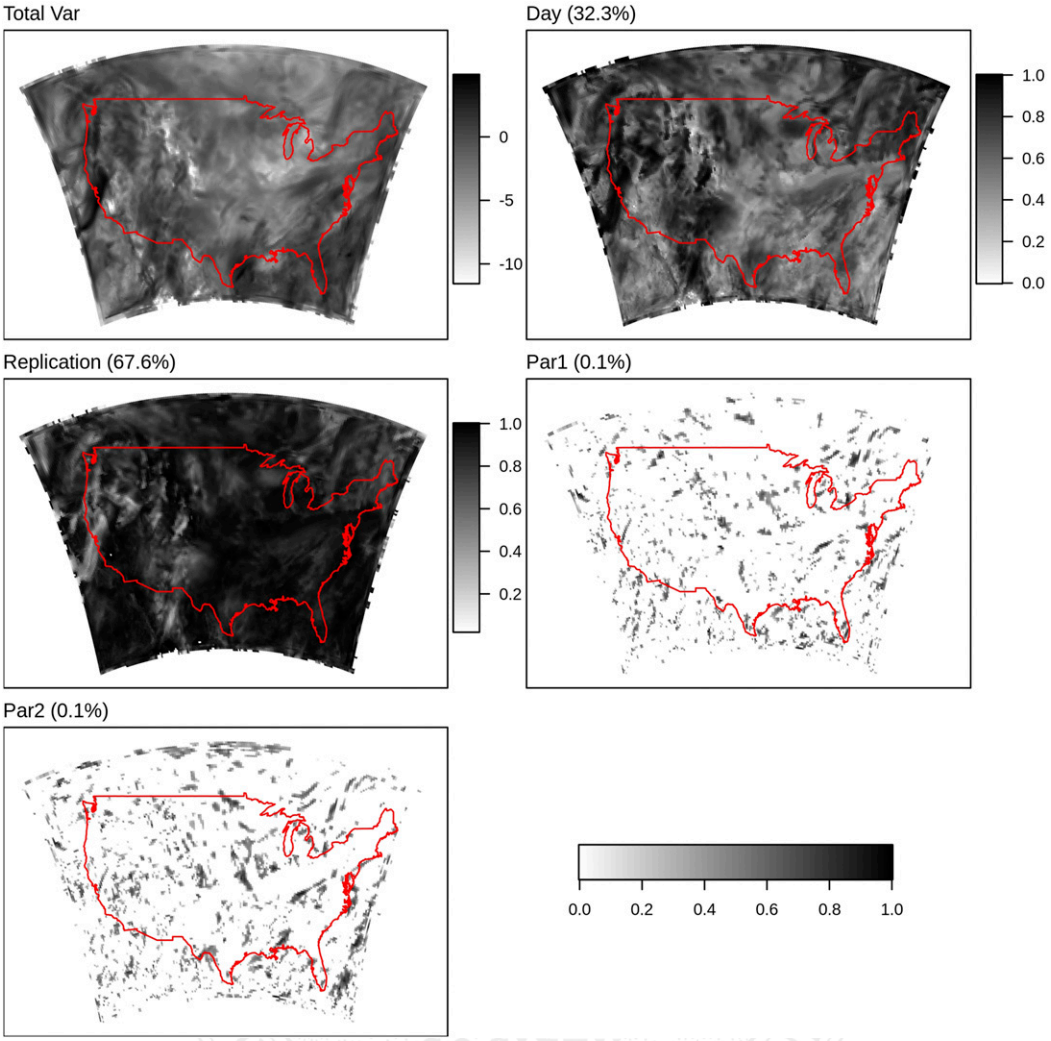

FIG. 3. (*Continued*)

existence of nontrivial spatial structure associated with the sensitivity of the various factors.

Possible explanations of the land–water contrast in Fig. 5 are as follows: Analyses of sea surface temperature (SST) are used as lower boundary conditions in our simulations. SST is a slow-varying variable at weather time scales, contributing to the reduced total (Total Var panel) and daily variances (Day panel in Fig. 5) over portions of the Pacific Ocean and Gulf of Mexico. The interaction of the lower atmosphere with the imposed lower boundary acts as a significant stabilizing influence that limits surface air temperature variability compared to temperatures over continental surfaces. This is particularly obvious over the Pacific Ocean portion of our domain. There, the proximity of imposed lateral boundary conditions at the inflow boundary (western edge of our model domain) leaves little time for perturbations to grow, particularly over the relatively short forecast horizons considered here. Both influences (from imposed lower and lateral boundaries) limit the extent to which variability in surface air temperature can develop.

This is in contrast to land surfaces, where two-way interactions between the atmosphere and land contribute to larger total and daily variances. But for replication, influences from land properties, terrain, and diurnal cycles dominate and force the replicated forecasts to be more similar, hence less sensitivity to replication over land. By contrast, the damped forcing over the ocean allows the variability of the replications to dominate, thereby explaining the reversal seen in Replication panel of Fig. 5. It is also noted that results over the Atlantic Ocean behave differently from those over the Pacific Ocean. Over the eastern part of the domain, contrasts

JOBNAME: MWR 00#00 2020 PAGE: 12 SESS: 8 OUTPUT: Thu Mar 26 16:56:34 2020 Total No. of Pages: 19
/ams/mwr/0/mwrD190321

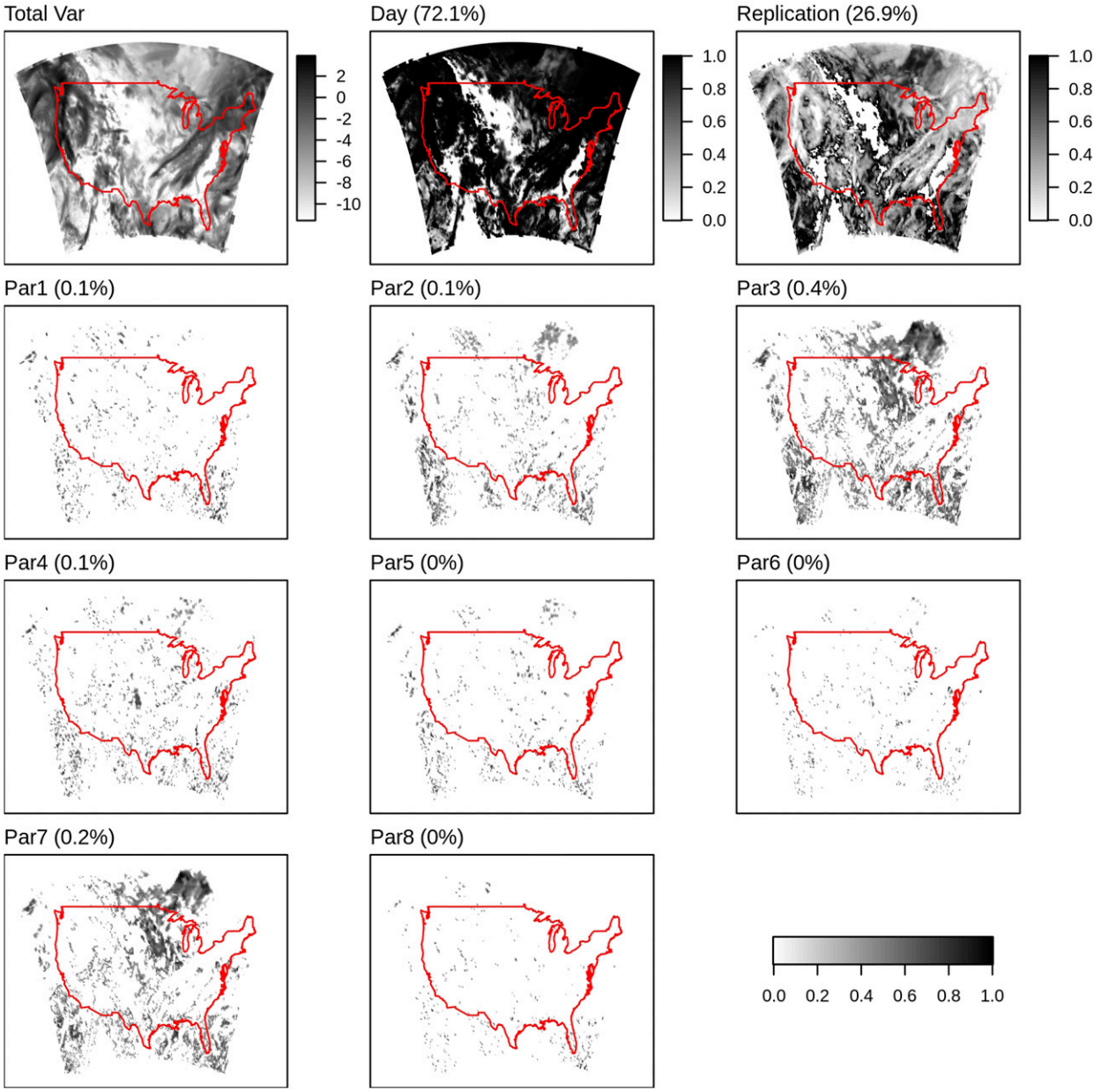

FIG. 4. (a) As in Fig. 3a, but for an FFD involving eight model parameters. Also, the quantity shown is the 95% LCB for ρ. (b) As in (a), but for 24-h forecasts of accumulated precipitation.

between land and ocean surfaces are not observed. Perturbations that develop over land and have time to amplify generally propagate over the ocean along the prevailing flow, contributing to the larger daily variability, at the expense of replication (see Fig. 5).

When the response is 3-h forecasts of 250-hPa wind speed, the results (not shown) suggest that parameters 1, 4, 6, and 8 have a similar and diffused spatial structure. The remaining parameters appear to have no effect at all. The results for 24-h forecasts, with the point estimates of ρ displayed, show highly nontrivial spatial patterns (Fig. 6). The absence of any topographic signature in these maps is not surprising, but the existence of any spatial structure at all is somewhat unexpected. The variability due to the Day factor still accounts for the majority of the total variance, and with similar spatial structure. The spatial structure of the Replication factor nearly complements that of the Day factor [i.e., where the sensitivity to daily variability is high (low), sensitivity to replication is low

JOBNAME: MWR 00#00 2020 PAGE: 13 SESS: 8 OUTPUT: Thu Mar 26 16:56:35 2020 Total No. of Pages: 19
/ams/mwr/0/mwrD190321

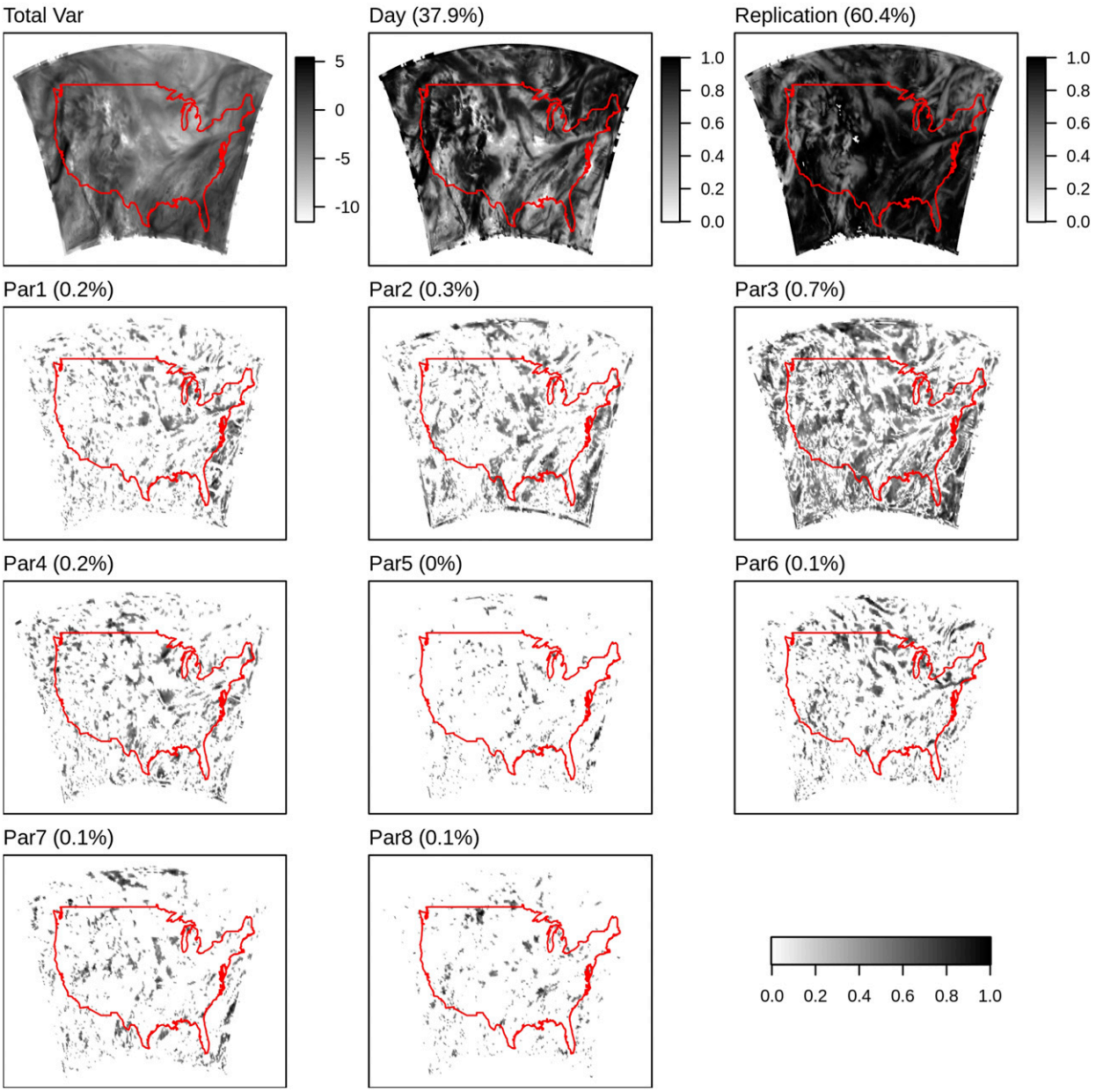

FIG. 4. (*Continued*)

(high)]. Par2 and Par3 have a nearly identical spatial structure, and that structure is different from those seen for the other model parameters.

As a final analysis, it is worth examining the domain-average sensitivity values (reported atop each panel in the figures above) because they display an interesting pattern. Table 4 summarizes the average ρ values for the Day factor, the Replication factor, and the total contribution from the model parameters; the GLSD results as well as the 3-h forecasts results (excluded above) are also included. Several patterns are noticeable:

- As one "moves" from 3- to 24-h forecasts, the contribution of the Day factor decreases, more so for precipitation than for wind speed and temperature.
- As one "moves" from 3- to 24-h forecasts, the contribution of the Replication factor increases, more so for precipitation than for wind speed and temperature.
- Of the three contributors to sensitivity—Day, Replication, and model parameters—the latter has the smallest effect.
- Of the two large contributors—Day and Replication—the former is larger, except for 24-h precipitation forecasts.

JOBNAME: MWR 00#00 2020 PAGE: 14 SESS: 8 OUTPUT: Thu Mar 26 16:56:35 2020 Total No. of Pages: 19
/ams/mwr/0/mwrD190321

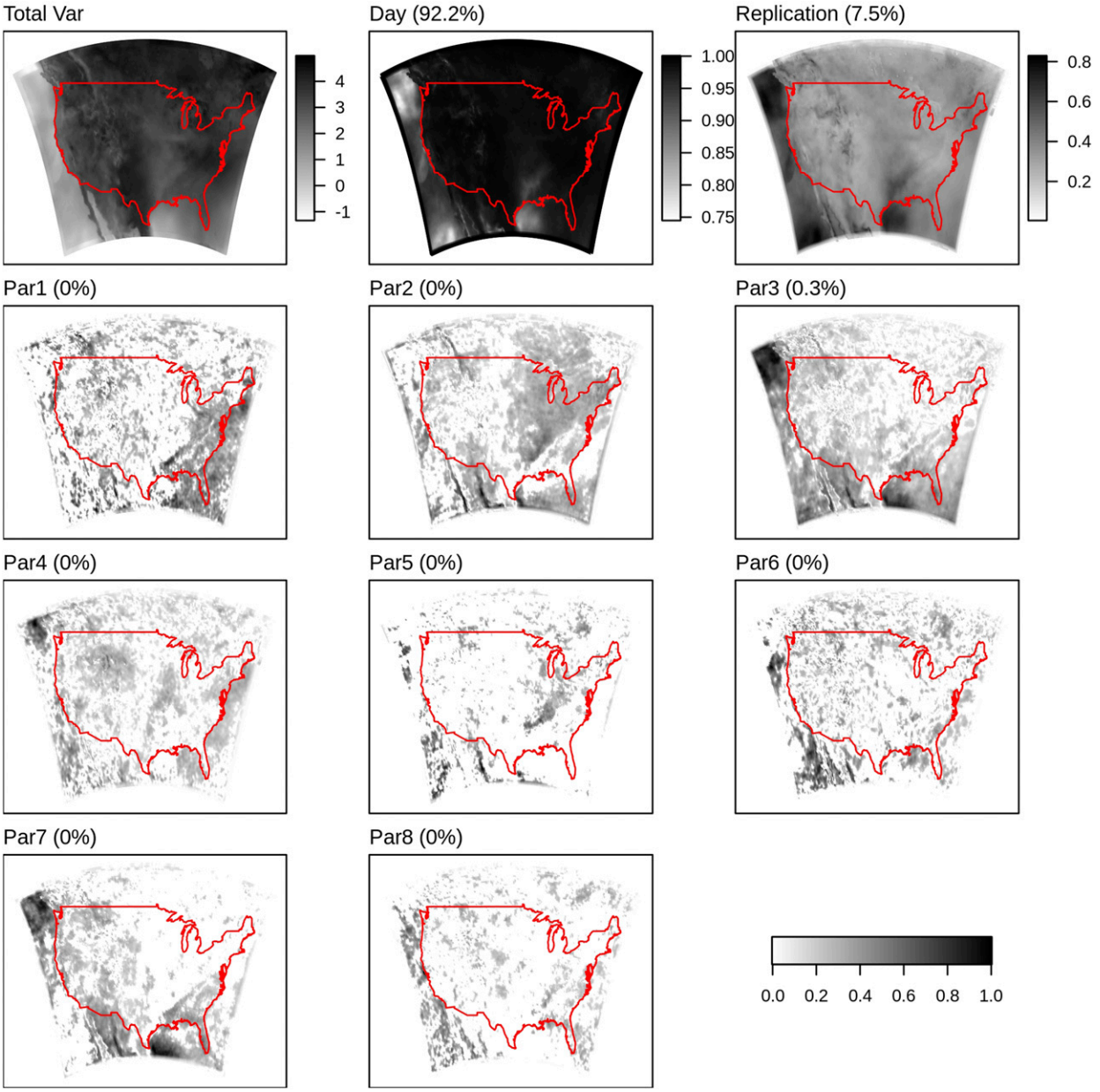

FIG. 5. As in Fig. 4b, but for 2-m temperature. Also, the quantity shown is the estimate of ρ itself, not the 95% LCB.

- FFD and LSD yield the same patterns.

All of these observations can be summarized as follows: for discrete forecast fields (e.g., precipitation) sensitivity with respect to daily variability decreases significantly with longer forecast times, while that due to replication increases significantly. But for continuous forecast fields such as temperature and wind speed, almost all of the sensitivity is due to daily variability, and it decreases only modestly for longer forecast times.

5. Conclusions and discussion

A methodology is developed for assessing the spatial structure of the sensitivity of forecasts with respect to noncontinuous factors. Random-effects models are employed to estimate the intraclass correlation (ρ) expressing the proportion of the total variability in forecasts that can be attributed to a given factor. Fractional factorial designs (FFD) and Graeco-Latin square designs (GLSD) are utilized to select the values of the factors. Maps of the estimated ρ values, as well as its

JOBNAME: MWR 00#00 2020 PAGE: 15 SESS: 8 OUTPUT: Thu Mar 26 16:56:36 2020 Total No. of Pages: 19
/ams/mwr/0/mwrD190321

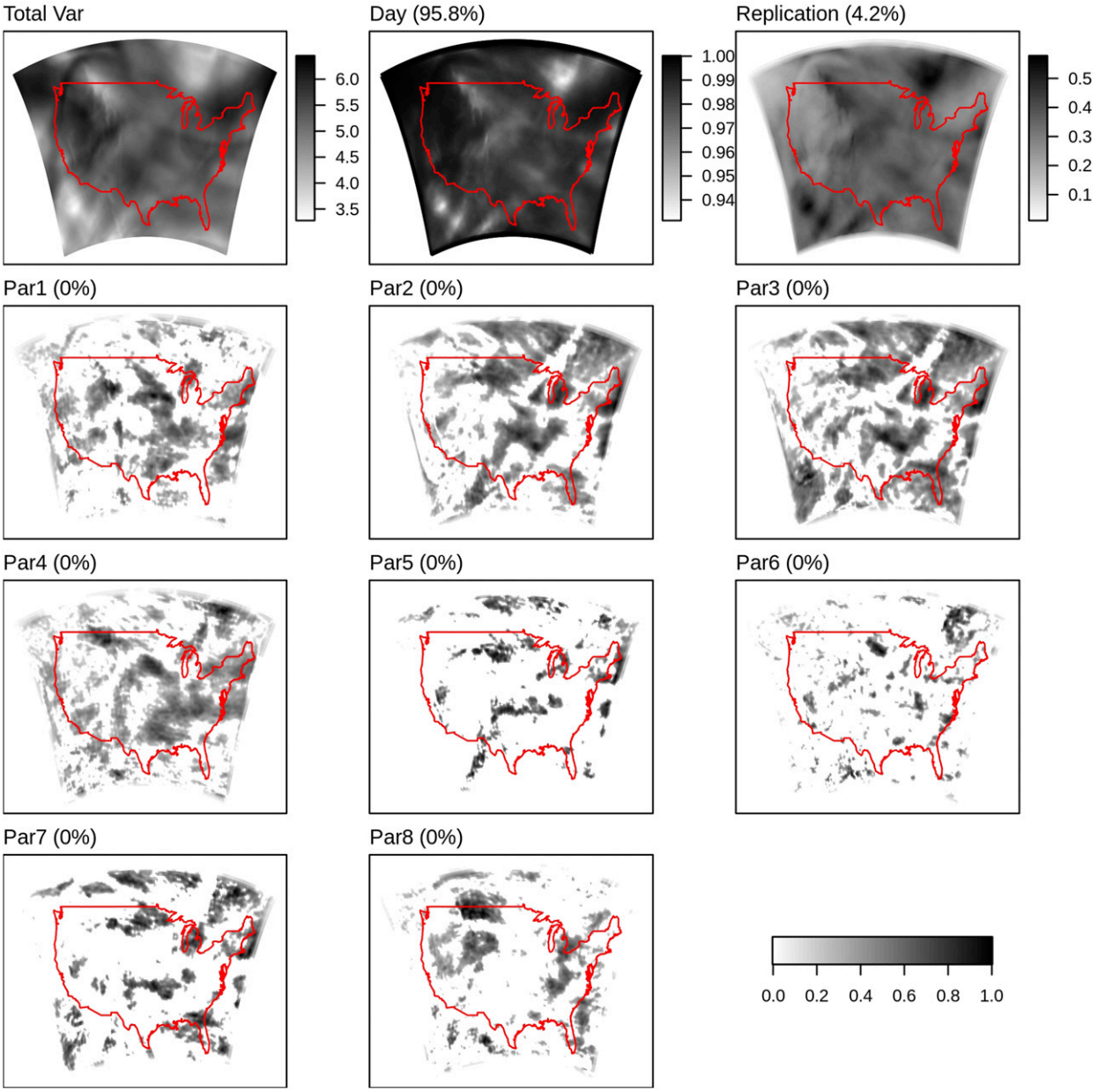

FIG. 6. As in Fig. 5, but for 250-hPa wind speed.

95% lower confidence bound, at each grid point, then provide for a visual display of the sensitivities.

The methodology is applied to 10 factors—1 representing daily variability, 1 representing replication, and 8 others corresponding to SKEBS/WRF Model parameters—for the purpose of studying the spatial structure of the effect of the factors on the forecasts of precipitation, temperature, and wind speed. The findings are complex, but they confirm previous findings that the model parameters have no appreciable effect in magnitude when compared with the effect of daily variability and replication. Furthermore, whereas the sensitivity of the temperature and wind speed forecasts with respect to the Day factor falls off for longer forecasts, that pattern is reversed for precipitation forecasts. More importantly, in spite of the small magnitude of the effects, the sensitivity associated with each of the factors examined here has a distinct spatial structure. A consequence of this finding is that one should not expect the effect of "fine-tuning" to be homogeneous across the forecast domain, especially for shorter lead times.

JOBNAME: MWR 00#00 2020 PAGE: 16 SESS: 8 OUTPUT: Thu Mar 26 16:56:37 2020 Total No. of Pages: 19
/ams/mwr/0/mwrD190321

TABLE 4. The average (across domain) percentage of total variability attributed to daily variability, Replication, and the model parameters in the 2^{k-p} (top) FFD and (bottom) LSD. The ±values are 1.96 times the standard error of the average ρ; the resulting interval is approximately a 95% confidence interval.

		Precipitation		2-m temperature		250-hPa wind speed	
		3 h	24 h	3 h	24 h	3 h	24 h
FFD	Day	72.1 ± 0.5	37.9 ± 0.6	99.0 ± 0.1	92.2 ± 0.3	99.9 ± 0.0	95.8 ± 0.2
	Replication	26.9 ± 0.5	60.4 ± 0.6	0.9 ± 0.1	7.5 ± 0.3	0.1 ± 0.0	4.2 ± 0.2
	Parameters	1.0 ± 0.1	1.7 ± 0.2	0.1 ± 0.0	0.3 ± 0.1	0.0 ± 0.0	0.0 ± 0.0
LSD	Day	74.5 ± 0.5	32.3 ± 0.6	98.3 ± 0.2	84.6 ± 0.4	99.6 ± 0.1	88.9 ± 0.4
	Replication	25.4 ± 0.5	67.6 ± 0.6	1.7 ± 0.2	15.3 ± 0.4	0.4 ± 0.1	11.1 ± 0.4
	Parameters	0.1 ± 0.0	0.1 ± 0.0	0.0 ± 0.0	0.1 ± 0.0	0.0 ± 0.0	0.0 ± 0.0

In examining maps of any quantity (one per grid point), it is important to acknowledge the lack of independence across grid points. For example, in applications that involve maps of p values, one introduces corrections that account for the spatial correlation between neighboring grid points (Elmore et al. 2006; Livezey and Chen 1983; Wilks 2006). These corrections are necessary for assessing whether an effect can be considered as statistically significant at a given significance level. Although such corrections were implemented in some of the sequel of papers to which this article belongs, the emphasis on the spatial structure of the sensitivities renders the corrections of secondary importance.[2] For instance, even though the ρ values in Fig. 6 may not meet any standard of significance (corrected or otherwise), the fact that Par2 and Par3 have nearly identical spatial structure is unlikely to be affected by corrections for multiple hypothesis testing across dependent grid points.

A discussion of the number of days analyzed in this study is in order. Here, a relatively small number of days are examined because of the pedagogical nature of the article, as reflected in Figs. 1 and 2. The reasons for the specific choice (i.e., 9) trace back to a requirement of GLSD (i.e., wherein all factors are required to have the same number of levels). As a result, the number of days is tied to the number of levels deemed adquate for the factors. If such a constraint is unrealistic in a given study, then a 2^{k-p} design should be considered instead. It is also important to point out that a relatively small number of days analyzed here do not immediately obviate the conclusions found here because a relatively small number of days will generally lead to relatively large p values, but it is unlikely to affect to spatial structure of the p values. In other words, as long as the 9 days examined here are representative of the population, there is no reason to doubt the conclusions reported here. Said differently, increasing the number of days in the dataset will likely lead to smaller p values, but the spatial structure of the p values is expected to be unaffected.

As mentioned previously, the discussions of spatial structure in this paper have been qualitative, and justified only because focus has been on visually comparing one sensitivity map with another. However, it is possible to quantify spatial structure. Indeed, SKEBS itself contains parameters that directly affect the spatial structure of the forecasts (Shutts 2005; Berner et al. 2011); these are the parameters of the cellular automaton algorithm that is employed in the SKEBS pattern generator. In principle, these parameters are independent of the model parameters studied in this article, and as such one cannot ask how the parameters of the cellular automaton algorithm affect the model parameters in Table 3. However, the spatial structure of the forecasts can be modeled in ways that summarize that spatial structure by a handful of quantities whose sensitivity with respect to the parameters in Table 3 can be assessed using the method described in this paper. For instance, consider the two images in Fig. 7. These images are F7
simulated Gaussian Random Fields (GRF) (Cressie 1993); such fields are parameterized with a number of quantities which directly control (and, therefore, quantify) the spatial structure. In addition to parameters that set the mean and variance of the entire field, there are also parameters that control the spatial extent of correlations, which in turn set the ''size'' of the typical ''object'' in the field; one such parameter is often denoted as scale, and Fig. 7 shows two examples of random Gaussian fields with different values of the scale parameter. It is worth pointing out that these

[2] In addressing the problem of multiple hypothesis testing, Marzban et al. (2018a) control the false discovery rate (Benjamini and Hochberg 1995), and Marzban et al. (2018b) use multivariate multiple regression and the Pillai's trace test (Fox et al. 2013; DelSole and Yang 2011) to account for spatial correlation.

JOBNAME: MWR 00#00 2020 PAGE: 17 SESS: 8 OUTPUT: Thu Mar 26 16:56:37 2020 Total No. of Pages: 19
/ams/mwr/0/mwrD190321

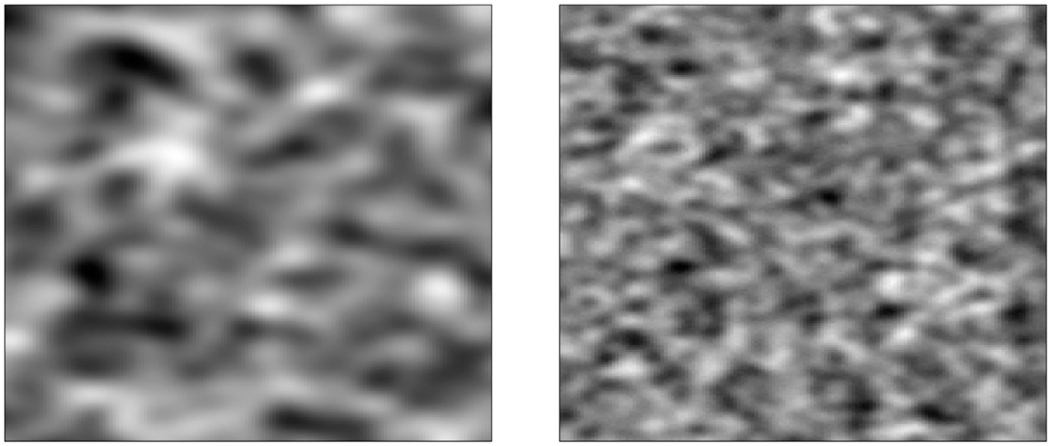

FIG. 7. Examples of simulated random Gaussian fields with the scale parameter set to (left) 10 and (right) 5.

examples do resemble many of the sensitivity maps shown in this paper, as well as the spatial patterns generated by the native SKEBS pattern generator [see e.g., Fig. 1 in Berner et al. (2011) and in Shutts (2005)]. As such, it is possible to apply the GRF model to the sensitivity maps generated here, and estimate the various parameters of the underlying Gaussian field. Then, by setting the response variable in Eq. (1) to one of these Gaussian parameters, one can assess the impact of the model parameters in Table 1 on the spatial struture of the forecast field. This quantification of spatial structure has, in fact, already been employed in assessing the quality of gridded forecast fields (Marzban and Sandgathe 2009), and it could be applied to the present sensitivity maps. Such an exercise would make for a worthwhile extension of the present analysis.

Acknowledgments. This work has received support from Office of Naval Research (N00014-12-G-0078 task 29) and National Science Foundation (AGS-1402895).

APPENDIX

Graeco-Latin Square and Fractional Factorial Designs

Consider an experiment involving four factors (A, B, C, D), each taking four values denoted A_i, B_i, C_i, and D_i, with $i = 1, \ldots, 4$. A full factorial design requires 4^4 runs, but it can be shown (Montgomery 2009) that if only the main effects (not interactions) are of interest, then only the specific 16 runs shown in Table A1 are sufficient. Such designs are called Graeco-Latin square designs (GLSD). Their defining characteristic is that every combination of the factor levels appears only once.

In a fractional factorial design (FFD) involving k binary factors, the number of runs is 2^k. However, there are certain designs that require only a fraction of those runs. These fractional deigns lead to aliasing (i.e., one can estimate only a linear combination of the effects). In some designs, however, the main effects are aliased only with high-order interaction effects. As such, under the assumption that the interactions are weak, one can use these designs to estimate the main effects. In one such design, of the 2^8 runs only the 16 shown in Table 2 are required. This design is called a 2^{8-4} (Resolution IV) design with generators ($\times 1$ $X2$ $X3$ $X7$), ($\times 1$ $X2$ $X4$ $X8$), ($\times 1$ $X3$ $X4$ $X6$), and ($\times 2$ $X3$ $X4$ $X5$). It can be shown that the alias structure of this design leads to the main effects to get aliased with three-way and higher interaction effects; all two-way interactions are aliased with each other, and none of them is aliased with the main effects (Montgomery 2009). In short, with this design one can reliably estimate the main effects (i.e., sensitivities) of eight factors with only 16 runs. This is the design used here for the FFD analysis.

TABLE A1. An example of an LSD involving four factors A, B, C, and D, each taking four values (denoted by the indices, 1, 2, 3, and 4).

	A_1	A_2	A_3	A_4
B_1	(C_1, D_1)	(C_2, D_2)	(C_3, D_3)	(C_4, D_4)
B_2	(C_2, D_4)	(C_1, D_3)	(C_4, D_2)	(C_3, D_1)
B_3	(C_3, D_2)	(C_4, D_1)	(C_1, D_4)	(C_2, D_3)
B_4	(C_4, D_3)	(C_3, D_4)	(C_2, D_1)	(C_1, D_2)

REFERENCES

Aires, F., P. Gentine, K. Findell, B. Lintner, and C. Kerr, 2014: Neural network–based sensitivity analysis of summertime convection over the continental United States. *J. Climate*, **27**, 1958–1979, https://doi.org/10.1175/JCLI-D-13-00161.1.

Ancell, B., and G. Hakim, 2007: Comparing adjoint- and ensemble-sensitivity analysis with applications to observation targeting. *Mon. Wea. Rev.*, **135**, 4117–4134, https://doi.org/10.1175/2007MWR1904.1.

Benjamini, Y., and Y. Hochberg, 1995: Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc.*, **57B**, 289–300, https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Berner, J., F. J. Doblas-Reyes, T. N. Palmer, G. Shutts, and A. Weisheimer, 2008: Impact of a quasi-stochastic cellular automaton backscatter scheme on the systematic error and seasonal prediction skill of a global climate model. *Philos. Trans. Roy. Soc. London*, **366A**, 2559–2577, https://doi.org/10.1098/rsta.2008.0033.

——, S. Y. Ha, J. P. Hacker, A. Fournier, and C. Snyder, 2011: Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multiphysics representations. *Mon. Wea. Rev.*, **139**, 1972–1995, https://doi.org/10.1175/2010MWR3595.1.

Bowman, K. P., J. Sacks, and Y.-F. Chang, 1993: Design and analysis of numerical experiments. *J. Atmos. Sci.*, **50**, 1267–1278, https://doi.org/10.1175/1520-0469(1993)050<1267:DAAONE>2.0.CO;2.

Cioppa, T., and T. Lucas, 2007: Efficient nearly orthogonal and space-filling Latin hypercubes. *Technometrics*, **49**, 45–55, https://doi.org/10.1198/004017006000000453.

Cressie, N. A. C., 1993: *Statistics for Spatial Data*. John Wiley and Sons, 900 pp.

DelSole, T., and X. Yang, 2011: Field significance of regression patterns. *J. Climate*, **24**, 5094–5107, https://doi.org/10.1175/2011JCLI4105.1.

Elmore, K. L., M. E. Baldwin, and D. M. Schultz, 2006: Field significance revisted: Spatial bias errors in forecasts as applied to the eta model. *Mon. Wea. Rev.*, **134**, 519–531, https://doi.org/10.1175/MWR3077.1.

Errico, R. M., 1997: What is an adjoint model? *Bull. Amer. Meteor. Soc.*, **78**, 2577–2591, https://doi.org/10.1175/1520-0477(1997)078<2577:WIAAM>2.0.CO;2.

Fasso, A., 2006: Sensitivity analysis for environmental models and monitoring networks. *Proc. iEMSs Third Biennial Meeting: Summit on Environmental Modelling and Software*, Burlington, VT, International Environmental Modelling and Software Society, http://www.iemss.org/iemss2006/sessions/all.html.

Fox, J., M. Friendly, and S. Weisberg, 2013: Hypothesis tests for multivariate linear models using the car package. *R J.*, **5**, 39–52, https://doi.org/10.32614/RJ-2013-004.

Hacker, J. P., C. Snyder, S.-Y. Ha, and M. Pocernich, 2011: Linear and non-linear response to parameter variations in a mesoscale model. *Tellus*, **63A**, 429–444, https://doi.org/10.1111/j.1600-0870.2010.00505.x.

Hodur, R. M., 1997: The Naval Research Laboratory's Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS). *Mon. Wea. Rev.*, **125**, 1414–1430, https://doi.org/10.1175/1520-0493(1997)125<1414:TNRLSC>2.0.CO;2.

Järvinen, H., M. Laine, A. Solonen, and H. Haario, 2012: Ensemble prediction and parameter estimation system: The concept. *Quart. J. Roy. Meteor. Soc.*, **138**, 281–288, https://doi.org/10.1002/qj.923.

Laine, M., A. Solonen, H. Haario, and H. Järvinen, 2012: Ensemble prediction and parameter estimation system: The method. *Quart. J. Roy. Meteor. Soc.*, **138**, 289–297, https://doi.org/10.1002/qj.922.

Leith, C. E., 1990: Stochastic backscatter in a subgrid-scale model: Plane shear mixing layer. *Phys. Fluids*, **2A**, 297–299, https://doi.org/10.1063/1.857779.

Li, X., N. Sudarsanam, and D. D. Frey, 2006: Regularities in data from factorial experiments. *Complexity*, **11**, 32–45, https://doi.org/10.1002/cplx.20123.

Livezey, R. E., and W. Y. Chen, 1983: Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.*, **111**, 46–59, https://doi.org/10.1175/1520-0493(1983)111<0046:SFSAID>2.0.CO;2.

Lucas, D. D., R. Klein, J. Tannahill, D. Ivanova, S. Brandon, D. Domyancic, and Y. Zhang, 2013: Failure analysis of parameter-induced simulation crashes in climate models. *Geosci. Model Dev.*, **6**, 1157–1171, https://doi.org/10.5194/gmd-6-1157-2013.

Marzban, C., 2013: Variance-based sensitivity analysis: An illustration on the Lorenz '63 model. *Mon. Wea. Rev.*, **141**, 4069–4079, https://doi.org/10.1175/MWR-D-13-00032.1.

——, and S. Sandgathe, 2009: Verification with variograms. *Wea. Forecasting*, **24**, 1102–1120, https://doi.org/10.1175/2009WAF2222122.1.

——, ——, J. D. Doyle, and N. C. Lederer, 2014: Variance-based sensitivity analysis: Preliminary results in COAMPS. *Mon. Wea. Rev.*, **142**, 2028–2042, https://doi.org/10.1175/MWR-D-13-00195.1.

——, C. Jones, N. Li, and S. Sandgathe, 2018a: On the effect of model parameters on forecast objects. *Geosci. Model Dev.*, **11**, 1577–1590, https://doi.org/10.5194/gmd-11-1577-2018.

——, X. D. S. Sandgathe, J. D. Doyle, Y. Jin, and N. C. Lederer, 2018b: Sensitivity analysis of the spatial structure of forecasts in mesoscale models: Continuous model parameters. *Mon. Wea. Rev.*, **146**, 967–983, https://doi.org/10.1175/MWR-D-17-0275.1.

——, R. Tardif, S. Sandgathe, and N. Hryniw, 2019: A methodology for sensitivity analysis of spatial features in forecasts: The stochastic kinetic energy backscatter scheme. *Meteor. Appl.*, **26**, 454–467, https://doi.org/10.1002/met.1775.

Mason, P. J., and D. J. Thomson, 1992: Stochastic backscatter in large-eddy simulations of boundary layers. *J. Fluid Mech.*, **242**, 51–78, https://doi.org/10.1017/S0022112092002271.

Montgomery, D. C., 2009: *Design and Analysis of Experiments*. 7th ed. Wiley & Sons, 656 pp.

Oakley, J. E., and A. O'Hagan, 2004: Probabilistic sensitivity analysis of complex models: A Bayesian approach. *J. Roy. Stat. Soc.*, **66B**, 751–769, https://doi.org/10.1111/j.1467-9868.2004.05304.x.

Ollinaho, P., H. Järvinen, P. Bauer, M. Laine, P. Bechtold, J. Susiluoto, and H. Haario, 2014: Optimization of NWP model closure parameters using total energy norm of forecast error as a target. *Geosci. Model Dev.*, **7**, 1889–1900, https://doi.org/10.5194/gmd-7-1889-2014.

Robock, A., and Coauthors, 2003: Evaluation of the North American land data assimilation system over the Southern Great Plains during warm season. *J. Geophys. Res.*, **108**, 8846, https://doi.org/10.1029/2002JD003245.

Roebber, P., 1989: The role of surface heat and moisture fluxes associated with large-scale ocean current meanders in maritime cyclogenesis. *Mon. Wea. Rev.*, **117**, 1676–1694, https://doi.org/10.1175/1520-0493(1989)117<1676:TROSHA>2.0.CO;2.

JOBNAME: MWR 00#00 2020 PAGE: 19 SESS: 8 OUTPUT: Thu Mar 26 16:56:38 2020 Total No. of Pages: 19
/ams/mwr/0/mwrD190321

——, and L. Bosart, 1998: The sensitivity of precipitation to circulation details. Part I: An analysis of regional analogs. *Mon. Wea. Rev.*, **126**, 437–455, https://doi.org/10.1175/1520-0493(1998)126<0437:TSOPTC>2.0.CO;2.

Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn, 1989: Design and analysis of computer experiments. *Stat. Sci.*, **4**, 409–423, https://doi.org/10.1214/ss/1177012413.

Safta, C., D. Ricciuto, K. Sargsyan, B. Debusschere, H. Najm, M. Williams, and P. Thornton, 2015: Global sensitivity analysis, probabilistic calibration, and predictive assessment for the data assimilation linked ecosystem carbon model. *Geosci. Model Dev.*, **8**, 1899–1918, https://doi.org/10.5194/gmd-8-1899-2015.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, M. Saisana, and S. Tarantola, 2008: *Global Sensitivity Analysis: The Primer*. Wiley Publishing, 304 pp.

——, P. Annoni, I. Azzini, F. Campolongo, M. Ratto, and S. Tarantola, 2010: Variance based sensitivity analysis of model output: Design and estimator for the total sensitivity index. *Comput. Phys. Commun.*, **181**, 259–270, https://doi.org/10.1016/j.cpc.2009.09.018.

Santner, T. J., B. J. Williams, and W. Notz, 2003: *The Design and Analysis of Computer Experiments*. Springer, 299 pp.

Shutts, G., 2005: A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Quart. J. Roy. Meteor. Soc.*, **131**, 3079–3102, https://doi.org/10.1256/qj.04.106.

Skamarock, W. C., and J. B. Klemp, 2008: A time-split nonhydrostatic atmospheric model for weather research and forecasting applications. *J. Comput. Phys.*, **227**, 3465–3485, https://doi.org/10.1016/j.jcp.2007.01.037.

Sobol', I. M., 1993: Sensitivity estimates for nonlinear mathematical models. *Math. Modell. Comput. Exp.*, **1**, 407–414.

Welch, W. J., R. J. Buck, J. Sacks, H. P. Wynn, T. J. Mitchell, and M. D. Morris, 1992: Screening, predicting, and computer experiments. *Technometrics*, **34**, 15–25, https://doi.org/10.2307/1269548.

Wilks, D. S., 2006: On "field significance" and the false discovery rate. *J. Appl. Meteor. Climatol.*, **45**, 1181–1189, https://doi.org/10.1175/JAM2404.1.

Zhao, J., and C. Tiede, 2011: Using a variance-based sensitivity analysis for analyzing the relation between measurements and unknown parameters of a physical model. *Nonlinear Processes Geophys.*, **18**, 269–276, https://doi.org/10.5194/npg-18-269-2011.