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

CAREN MARZBAN

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

XIAOCHUAN DU

Department of Statistics, University of Washington, Seattle, Washington

SCOTT SANDGATHE

Applied Physics Laboratory, University of Washington, Seattle, Washington

JAMES D. DOYLE AND YI JIN

Naval Research Laboratory, Monterey, California

NICHOLAS C. LEDERER

The Boeing Company, Applied Mathematics, Seattle, Washington

(Manuscript received 25 September 2017, in final form 22 January 2018)

ABSTRACT

A methodology is proposed for examining the effect of model parameters (assumed to be continuous) on the spatial structure of forecasts. The methodology involves several statistical methods of sampling and inference to assure the sensitivity results are statistically sound. Specifically, Latin hypercube sampling is employed to vary the model parameters, and multivariate multiple regression is used to account for spatial correlations in assessing the sensitivities. The end product is a geographic ''map'' of p values for each model parameter, allowing one to display and examine the spatial structure of the sensitivity. As an illustration, the effect of 11 model parameters in a mesoscale model on forecasts of convective and grid-scale precipitation, surface air temperature, and water vapor is studied. A number of spatial patterns in sensitivity are found. For example, a parameter that controls the fraction of available convective clouds and precipitation fed back to the grid scale influences precipitation forecasts mostly over the southeastern region of the domain; another parameter that modifies the surface fluxes distinguishes between precipitation forecasts over land and over water. The sensitivity of surface air temperature and water vapor forecasts also has distinct spatial patterns, with the specific pattern depending on the model parameter. Among the 11 parameters examined, there is one (an autoconversion factor in the microphysics) that appears to have no influence in any region and on any of the forecast quantities.

1. Introduction

The algorithms in numerical weather prediction models contain numerous parameters—hereafter, ''model parameters''—whose values and/or ranges are generally established by field or laboratory experiments or on theoretical grounds (Stensrud 2007). The effect of these parameters on forecasts is important, but often complex and difficult to establish. Two techniques that are often used to address the influence of model parameters on forecasts are sensitivity analysis and model tuning (also called fine-tuning or calibration). This article proposes a sensitivity analysis method aimed at better understanding the effect of model parameters on the spatial structure of forecasts. The model used to illustrate

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

DOI: 10.1175/MWR-D-17-0275.1

the method is Coupled Ocean–Atmosphere Mesoscale Prediction System (COAMPS),[1] and the forecast quantities are precipitation (convective and resolved grid scale), surface air temperature, and water vapor.

It is important to distinguish between the two approaches because in spite of their similarities, their focuses are different. The immediate goal of model tuning is the objective improvement of forecasts (Duan et al. 2006; Hourdin et al. 2017; Voudouri et al. 2017). While that effort does naturally involve assessing the sensitivity of the forecasts with respect to model parameters, the objective improvement of forecasts is only an indirect and long-term objective in sensitivity analysis. The main focus of sensitivity analysis is to establish the conditions under which parameters have an effect at all, and then to quantify the magnitude and statistical significance of the effects (Saltelli et al. 2004, 2010; Warner 2011). The difference between the two methods can be seen in the fact that model tuning relies on observations in order to guide the tuning of the model, whereas in sensitivity analysis, one is interested mostly in the effect of model parameters on forecasts only. As such, model tuning can be thought of as a sensitivity analysis of forecast errors.

The covariates (i.e., the independent variables) in a sensitivity analysis or model tuning are not always model parameters; instead, one may be interested in the effect of initial conditions, the impact of observations and/or their location, the effect of the various members of an ensemble, sensitivity with respect to a unit change in a state variable, or combinations of all of the above (Ancell and Hakim 2007; Daescu and Langland 2013; Davis and Emanuel 1991; Gombos and Hansen 2008; Hacker et al. 2011; Järvinen et al. 2012; Laine et al. 2012; Ollinaho et al. 2014; Torn and Hakim 2008; Weisman et al. 2015). Based on the above definitions, these examples broadly—but by no means categorically—fall under the umbrella of model tuning. Examples of studies that place more emphasis on sensitivity analysis—but not to the exclusion of model tuning—are Adlerman and Droegemeier (2002), Boyle et al. (2015), Crook (1996), Gómez-Navarro et al. (2015), Houston and Niyogi (2007), Marzban et al. (2014), Mölders et al. (1995), Qian et al. (2015), Robock et al. (2003), Roebber (1989), Roebber and Bosart (1998), Schumann and Roebber (2010), and Zhao and Tiede (2011).

As mentioned above, the immediate goal of sensitivity analysis is to expose the conditions under which the parameters have an effect on the forecasts and to assess the strength of that effect. In Marzban et al. (2014), the effects of 11 model parameters (Table 1) on the domain average and center of gravity of precipitation are examined. Here, in order to provide a visual display of the spatial structure of sensitivity across the entire domain, a measure of sensitivity is produced at each and every grid point.

On a spatial/gridded field, estimates of sensitivity are affected by dependency (spatial correlation) between grid points. Such correlations affect both the point estimate of sensitivity, as well as tests of statistical significance. Another complicating factor is the multiplicity of tests arising from testing either multiple model parameters at a given grid point or across multiple grid points (Wilks 2006, 2011). It is well known that multiple hypothesis testing leads to an increase in type I errors: for example, suggesting that a model parameter has a significant effect on the forecasts at a given grid point, when in fact it does not. As such, it is important to avoid, minimize the number of, or take steps to account for spatial correlations and the multiplicity of hypothesis tests (Benjamini and Hochberg 1995; Bretz et al. 2010; Dmitrienko et al. 2009; Noble 2009; Rosenblatt 2013; Wilks 2011).

The present paper puts forth a methodology that allows one to address all of these issues. The methodology consists of several well-known techniques from the field of experimental design. Latin hypercube sampling is employed to select the values of the model parameters across which they are varied. Multivariate multiple regression (MMR), a generalization of multiple regression to the case where multiple predictors and multiple responses are present, is used to model the relationship between 11 model parameters and forecasts at several grid points, simultaneously. The multivariate (i.e., several responses) nature of MMR allows one to account for spatial correlations and the multiplicity of those tests pertaining to multiple model parameters and multiple responses. In the proposed approach, the problems associated with multiple hypothesis tests across the multitude of grid points are avoided completely, because at no stage are the p values compared with a significance level of any kind for the purpose of hypothesis testing; the magnitude of the p values is sufficient to provide a visual assessment of the significance and the magnitude of the sensitivities.

These steps lead to spatially corrected p values that are displayed as a geographic ‘‘map’’ reflecting the spatial structure of the sensitivities, first across all model parameters and then for each model parameter. These maps are produced for each of the 40 days/forecasts in the dataset examined here. Although such daily sensitivity maps can be useful for some users, the question of how model parameters affect forecasts requires an assessment of sensitivity that is independent of day/time. To that end, a final statistical test is applied to produce a

[1] COAMPS is a registered trademark of the Naval Research Laboratory.

TABLE 1. The 11 parameters studied in this paper. Also shown are the default values and the range over which they are varied.

ID	Name (unit)	Description	Default	Range
1	delt2KF (°C)	Temperature increment at the LCL for KF trigger	0	−2, 2
2	cloudrad (m)	Cloud radius factor in KF	1500	500, 3000
3	prcpfrac	Fraction of available precipitation in KF, fed back to the grid scale	0.5	0, 1
4	mixlen	Linear factor that multiplies the mixing length within the PBL	1.0	0.5, 1.5
5	sfcflx	Linear factor that modifies the surface fluxes	1.0	0.5, 1.5
6	wfctKF	Linear factor for the vertical velocity (grid scale) used by KF trigger	1.0	0.5, 1.5
7	delt1KF (°C)	Another method to perturb the temperature at the LCL in KF	0	−2, 2
8	autocon1 ($\mathrm{kg\,m^{-3}\,s^{-1}}$)	Autoconversion factors for the microphysics	0.001	1×10^{-4}, 1×10^{-2}
9	autocon2 ($\mathrm{kg\,m^{-3}\,s^{-1}}$)	Autoconversion factors for the microphysics	4×10^{-4}	4×10^{-5}, 4×10^{-3}
10	rainsi ($\mathrm{m^{-1}}$)	Microphysics slope intercept parameter for rain	8.0×10^{6}	8.0×10^{5}, 8.0×10^{7}
11	snowsi ($\mathrm{m^{-1}}$)	Microphysics slope intercept parameter for snow	2.0×10^{7}	2.0×10^{6}, 2.0×10^{8}

single map of p values that displays the spatial structure of sensitivities aggregated across days.

2. Method

a. Data

The methodology developed here requires running the forecasting model (here, COAMPS) for different values of the model parameters (and for different days/forecasts). The model output—henceforth, "data"—generated in this way is then used for the sensitivity analysis. Given the large computational requirements of running forecasting models, the number of times one can run them is an important practical consideration. Methods of experimental design (Montgomery 2009) can be used to both minimize the number of runs and to select the specific values of the model parameters for each run. The experimental design of this study is similar to that of Marzban et al. (2014), and so, it is described here only briefly. Only the atmospheric portion of COAMPS (Hodur 1997) is used. For 40 days, at approximately 3-day intervals between 16 February and 2 July 2009, COAMPS 24-h forecasts are generated for four meteorological quantities: 24-h accumulated convective and grid-scale precipitation, surface air temperature, and water vapor. For each date, a 99-member ensemble is produced by varying 11 model parameters; these parameters are shown in Table 1, and the reasons for their selection are explained by Holt et al. (2011).

The specific values of the parameters are obtained by taking a Latin hypercube sample (LHS) of size 99 from the 11-dimensional space of the parameters. Technically, by virtue of being an analysis of data produced by a computer model, the sensitivity analysis performed here is an instance of a computer experiment, wherein the sampling method of choice is LHS (Cacuci et al. 2005; Saltelli et al. 2004, 2010; Bowman et al. 1993; Fang et al. 2006; Sacks et al. 1989; Santner et al. 2003). This sampling scheme is designed to assure that no two of the 99 points have the same value for any of the 11 parameters. It can be shown that this property leads to estimates that are generally more precise than random sampling or varying the model parameters one at a time (Cioppa and Lucas 2007; Hacker et al. 2011; McKay et al. 1979; Marzban 2013; Marzban et al. 2014; Qian et al. 2015).

Here, the number of runs is set to 99, mostly based on trial and error; it is confirmed that fewer runs (20 and 50) produce similar results, although, not surprisingly, with lower statistical significance. The choice of the number of runs is also weighed against the time required for each of the COAMPS runs. To expedite the runs, a low-resolution configuration is used. Specifically, the size of the domain is 45 × 72 with 81-km grid spacing. Although this spacing may be too coarse to be useful for practical purposes, as shown below, it is sufficiently fine to demonstrate the methodology and to reveal many nontrivial and statistically significant spatial structures. The resolution of the model and the number of runs are two quantities that affect computational effort, and so their values depend on the computational resources available to users of this methodology. Regardless of computational resources, some trial and error is recommended in order to test the sensitivity of the final results on these two quantities.

The LHS is designed for situations where the variables being sampled are continuous. In the present application, all of the model parameters examined are continuous. The case where model parameters are discrete or categorical requires a different class of sampling schemes; that methodology will be considered elsewhere. The LHS is not the only sampling procedure with desirable properties; Qian et al. (2015) use LHS and an alternative known as quasi–Monte Carlo sampling. Further comparisons of the two sampling methods have been performed in Kucherenko et al. (2015).

b. Multivariate multiple regression

The statistical model used for assessing sensitivity and its statistical significance is MMR. The main reason for this choice is the manner in which MMR allows one to account for spatial correlations (DelSole and Yang 2011). Although other methods exist that take spatial correlations into account when performing inference (Douglas et al. 2000; Elmore et al. 2006; Wilks 1997), the MMR approach is more natural in the present application because the sensitivity analysis is done within a regression framework already.

The terms ''multivariate'' and ''multiple'' in MMR refer to several response and predictor variables, respectively. The benefits of an MMR model over several multiple regression models (one per response variable) are similar to the advantages of a single multiple regression model over several simple regression models (one for each predictor). Consider the latter comparison; it can be shown that if the predictors are completely uncorrelated, then multiple regression is completely equivalent to several simple regressions developed on each of the predictors. However, in the presence of any correlation between the predictors, the least squares regression coefficients in multiple regression are different from those in simple regression, and only the former are correct (i.e., unbiased estimates of the true population regression coefficients). In this sense, multiple regression ''accounts for'' the correlations between predictors.

Similarly, given several response (and several predictor) variables, if the responses are completely uncorrelated, then MMR has no advantages over several multiple regression models (one per response). But if there exists any correlation between the response variables, then the least squares estimates of the regression coefficients in multiple regression models are incorrect; only those obtained from MMR are unbiased estimates of the true/population values. In this sense, MMR accounts for correlations between response variables, as well as correlations between the predictors.

In the present application, the predictors in MMR are the 11 model parameters. To be able to compare the effect of the model parameters, they are all standardized (to have a mean and variance of 0 and 1, respectively) at each grid point. The response variables are the forecasts (e.g., air temperature) at multiple grid points. Here, the number of response variables is set to nine, corresponding to the grid points in a 3×3 window. Such a window corresponds to a square of size 243 km (3×81 km) on the side. In other words, it is assumed that the spatial correlation extends to that distance scale. This window size is selected to be sufficiently large to account for some spatial correlation, but sufficiently small to allow nontrivial spatial structure to be seen across the entire forecast domain. Henceforth, this window will be referred to as the MMR window.

The multivariate nature of MMR allows a variety of statistical tests. At the simplest level, one can test whether any of the predictors (model parameters) have an effect on any of the response variables (grid points in the MMR window). Hypotheses containing the word ''any,'' or equivalently, the phrase ''at least 1,'' are commonplace in statistics, and there exists a number of standard tests for them; such tests—often called ''omnibus''—lead to a single p value, and as such, avoid the complexities associated with multiple hypothesis testing arising from the multitude of predictor and response variables (Montgomery 2009; Wilks 2011). Only if the omnibus test is significant does one proceed to test the effect of each predictor; otherwise, there is no justification to test each predictor. This two-step testing procedure—an omnibus test followed by a sequence of more diagnostic tests—is a standard method for taming the adverse effects of multiple hypothesis testing (Montgomery 2009). The test used here is called the Pillai's trace test (Fox et al. 2013), and it leads to a single p-value map for each forecast day.

Although such daily maps of sensitivity may be useful for users who may expect the sensitivities to depend on certain types of weather, assessing sensitivity independently of forecast day is arguably more useful. To that end, the daily maps of p values are aggregated into a single p-value map; see the next subsection for an explanation of how this aggregation is performed. In short, MMR allows one to generate a single map of p values displaying the spatial structure of sensitivities across all model parameters and days. The utility of such a map is derived from the underlying omnibus test. As such, only if/when this map reveals a distinct spatial structure does one proceed to assess the contribution from each model parameter separately. This map is one of the main outputs of the proposed methodology.

The other main output is a number of p-value maps for examining the contribution of each model parameter separately. MMR allows for such a test as well; in this case, each p value assesses whether a given model parameter has an effect on any of the grid points in the MMR window. Given that the test is done within the MMR framework (i.e., with multiple responses and predictors), the resulting p values continue to take into account spatial correlations across the grid points and correlations between the model parameters.

In the proposed methodology, complexities associated with the multiplicity of predictors, responses, and grid points are addressed in different ways. First, the p values computed in MMR are penalized for larger number of predictors and/or responses; these numbers appear explicitly in the degrees of freedom associated with the test and in a way that generally increases the p value when the numbers are large (Montgomery 2009). This ''correction'' does not directly address the problems associated with multiple hypothesis testing because those problems arise

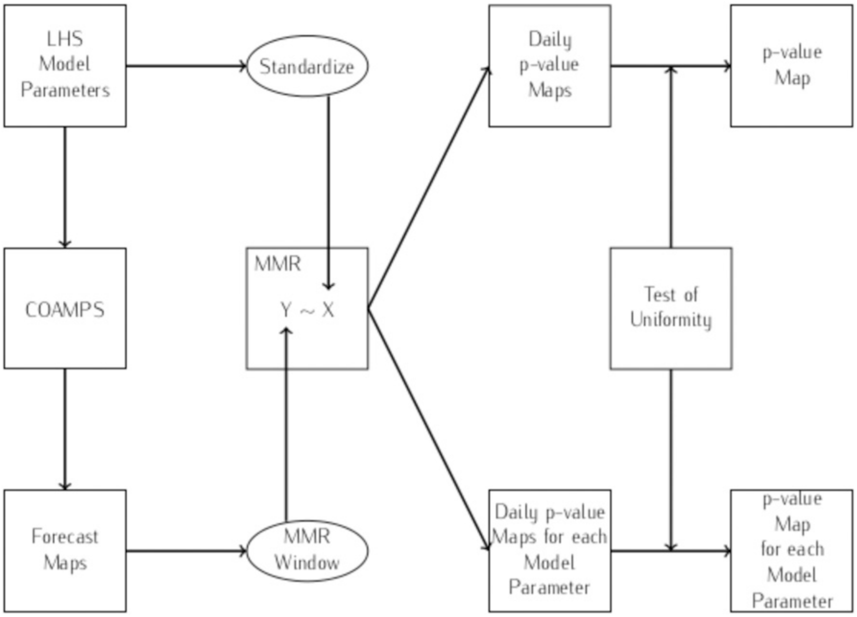

FIG. 1. A flowchart of the proposed methodology. "Map" denotes a spatial, gridded field. See text for explanation.

when one compares p values with a fixed number (e.g., significance level) for the purpose of rejecting or not rejecting a null hypothesis. Here, the problems associated with multiple hypothesis testing across the grid points are avoided altogether because at no stage are any of the p values employed for hypothesis testing.

The flowchart in Fig. 1 shows all of the elements of the proposed methodology. One begins with an LHS taken from the space of model parameters (top-left corner of Fig. 1). These parameters are fed into the mesoscale model (here, COAMPS), and forecast spatial/gridded fields—maps, for short—are produced. The model parameters are standardized and used as predictors in MMR, while the response variables for MMR are supplied by the values of forecasts inside the MMR window. The least squares estimates of MMR are subjected to two types of tests: 1) a test of whether any of the model parameters have an effect on any of the grid points in the MMR window and 2) a test of whether a given model parameter has an effect on any of the grid points in the MMR window. There exists one such p-value map per day. A test of uniformity, discussed in the next subsection, is applied to aggregate the daily p-value maps.

c. Why *p* values?

One may wonder why the proposed methodology places so much emphasis on producing p values when, ultimately, no hypothesis tests are performed at all.

There are three reasons for using p values to assess sensitivity. First, p values have a special property that can be used for a variety of purposes: they have a uniform distribution under the null hypothesis (Montgomery 2009). In the present application, this important property is employed to aggregate the 40 daily p-value maps into a single map of p values.

This aggregation involves the temporal component of the data, and because the 40 cases are selected to be 3 days apart, one may assume that they are reasonably independent. Under the null hypothesis that the forecast at a grid point is unaffected by any of the model parameters, the 40 p values at a given grid point ought to have a uniform distribution. Then, any violation of uniformity suggests that the parameter in question has an effect on the forecast at that grid point. Here, a chi-squared and a Kolmogorov–Smirnov test of uniformity have been performed, although only the results of the former are shown; the latter produced similar results. Again, the p values are displayed as a map.

The second reason for using p values is that although statistical significance and the magnitude of an effect are distinct concepts, the relationship between the magnitude of p values and the magnitude of the regression coefficients they test is monotonic. As such, a p value can be used not only for assessing statistical significance, but also to convey information about the magnitude of the sensitivity effect.

The proposed methodology has been developed to avoid multiple hypothesis testing, but not to its exclusion. The fact that the proposed methodology produces

p values at each grid point readily allows for multiple testing because all of the procedures developed for multiple hypothesis testing begin with a set of ''raw'' p values (such as those in the p-value maps) and then correct them for the multiplicity of tests (Benjamini and Hochberg 1995; Bretz et al. 2010; Dmitrienko et al. 2009; Rosenblatt 2013; Wilks 2011). It is this feature of multiple hypothesis testing procedures that constitutes the third reason for using p-value maps for displaying sensitivity. In other words, p values are used here because they can be readily corrected for multiple hypothesis testing; other measures of sensitivity (e.g., the regression coefficients) do not have this property.

However, the methodology developed here does not include the necessary corrections for multiple hypothesis testing because the corrections depend on the specification of an error rate to be controlled, and the choice of that error rate is highly user- or problem-dependent (Rosenblatt 2013). There exists a large number of error rates, similar to the wide variety of verification measures, but two common choices are the family-wise error rate, defined as the probability of at least one type I error, and the false discovery rate, which is the expected proportion of type I errors among all the tests that leads to the rejection of the null hypothesis. The ultimate decision to reject or not reject a null hypothesis depends on the choice of the error rate. It is worth mentioning that these corrections, too, utilize the uniformity of p values under the null hypothesis.

3. Application

To set the stage for the sensitivity analysis, for each of the four forecast quantities, Fig. 2 shows the average across 40 days (forecasts) and 99 parameter values (ensemble members). It is evident that the southeast region, off the coast of Florida, receives the most convective precipitation across the domain (Fig. 2a). Grid-scale precipitation (Fig. 2b) is far less structured, with a slightly higher amplitude off the eastern coast and across the Gulf states. The Kain–Fritsch (KF) scheme (Kain and Fritsch 1993) is sensitive to grid spacing, and it is expected that at this coarse resolution (81 km), the convective parameterization plays an important role in producing precipitation (Gallus 1999). Surface air temperature (Fig. 2c) displays the expected gradient with respect to latitude and with cooler temperatures over the Rockies and the Appalachian Mountains. Water vapor near the surface (Fig. 2d) displays a similar pattern to surface air temperature, with the exception that the low-level dry air extends farther south, into Mexico.

a. *Nonspatial sensitivity*

To make contact with work done previously (Marzban et al. 2014), first, the sensitivity of the domain mean of the forecast quantities is computed for each of the 11 model parameters. The data used for estimating these regression coefficients are the 99 cases corresponding to 99 samples taken from the 11 model parameters, and so there exists a regression coefficient for each model parameter and each of the 40 days in the dataset. Figure 3 shows these regression coefficients with the box plots displaying their variability across the 40 days.

If a boxplot is entirely above or below the horizontal line at $y = 0$, then one may conclude that the corresponding parameter has a consistent and significant effect on the forecast for all 40 days examined here. A relatively small overlap of the boxplot with the horizontal line implies that the parameter has a nontrivial effect, but the effect is weaker. And if a boxplot is nearly centered on the horizontal line, then one may conclude that the corresponding parameter has no effect on the forecast. Such plots are useful because they convey information about both statistical significance and the magnitude of the effect via, respectively, the spread of the boxplots and the overall location of the boxplots relative to the horizontal line at $y = 0$.

It can be seen that parameters 1, 3, and 7 have significant effects on both types of precipitation (Figs. 3a,b), though their effect on grid-scale precipitation is opposite to that on convective precipitation. Some explanation for this finding can be offered: parameter 3 controls the fraction of available precipitation fed back from the convection parameterization to the microphysics for grid-scale precipitation (Table 1). Alternatively, it may be considered as cloud detrainment primarily occurring at upper levels, where cloud ice usually forms; the condensate is then passed to the grid-scale microphysics. This connection represents the interaction between the grid-scale-resolved clouds and parameterized convection. When a large fraction of the available precipitation is provided to the grid-scale microphysics, there is less chance for the model to generate convective precipitation, and hence, a negative sensitivity to this parameter is expected (Fig. 3a). On the other hand, the grid-scale precipitation increases when more moisture becomes available, consistent with its positive sensitivity to parameter 3 (Fig. 3b). The effects of parameters 1 and 7 [controlling the temperature increment at the lifted condensation level (LCL) for the Kain–Fritsch cumulus parameterization] may be explained as follows: adding positive air temperature perturbations at the LCL may cause the air parcel to become more buoyant, rising faster, which is conducive to convective precipitation and is reflected in the positive sensitivity in Fig. 3a. The negative sensitivity of parameter 7 in Fig. 3b suggests that grid-scale precipitation decreases with the positive temperature

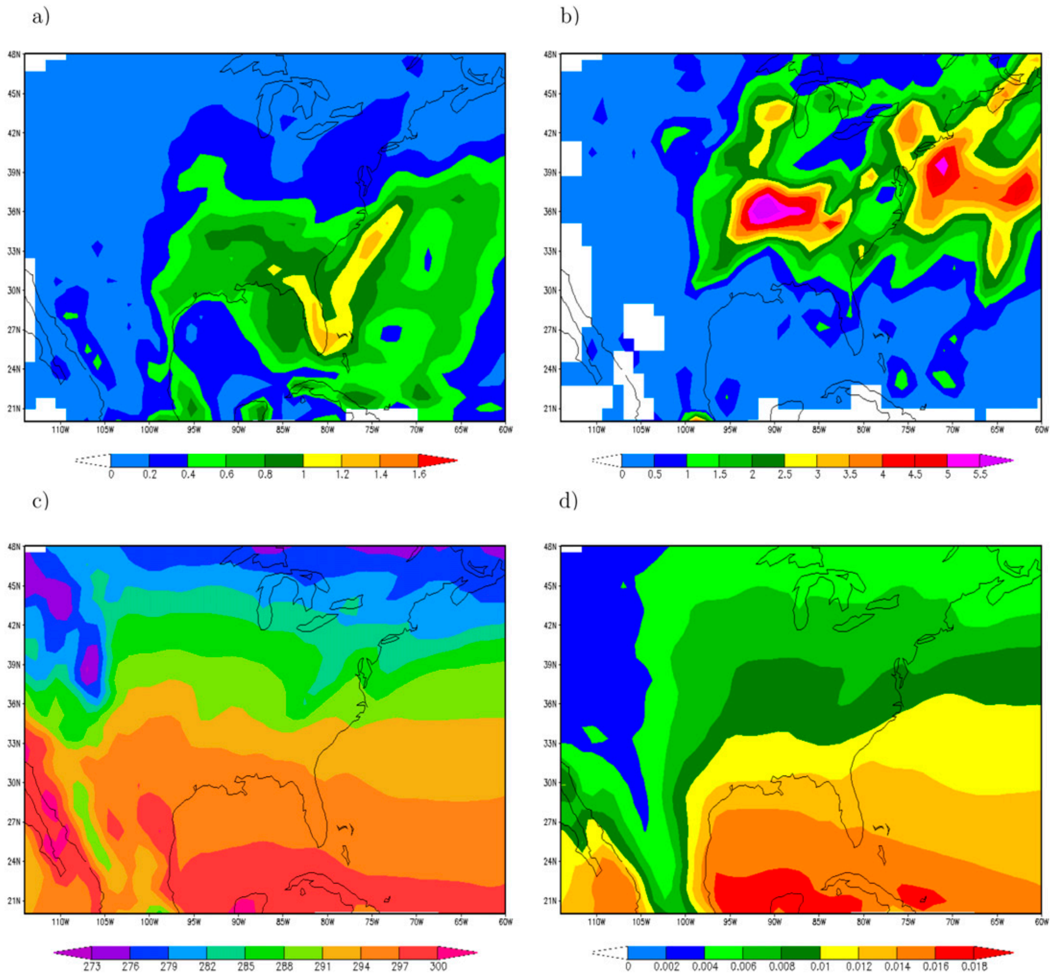

FIG. 2. The average (across 40 days and 99 parameter values) of (a) convective precipitation (mm), (b) grid-scale precipitation (mm), (c) surface temperature (K), and (d) water vapor (kg kg^{-1}). The white patches appearing in (a),(b) correspond to grid points at which no precipitation occurs across the 40 days and the 99 model parameter values. In the assignment of colors to precipitation values, a log-transformation has been applied in order to visually enhance the spatial structure of the forecasts.

perturbation at the LCL. In short, for any given amount of moisture available for precipitation, the cumulus and microphysics schemes represent two competing processes for precipitation production. As such, it is not surprising that the microphysics would generate less grid-sale precipitation as the cumulus scheme becomes more active for producing convective precipitation with the larger positive temperature perturbation at the LCL.

Another difference between Figs. 3a and 3b is in the effect of parameter 9, which represents the threshold beyond which clouds are converted into rain in the microphysics scheme (Rutledge and Hobbs 1983); it has a positive impact on grid-scale precipitation but no effect on convective precipitation. One possible explanation is that while the grid-scale precipitation production is delayed due to the increased value of parameter 9, the clouds have more time to develop, and the model actually produces more grid-scale precipitation over the accumulation period.

Surface air temperature (Fig. 3c) is affected by many of the parameters (to varying degrees, both positively and negatively), with the exception of parameters 3, 6, 8, 10, and 11, which appear to have no consistent effect across the 40 days. Interestingly, in regards to water vapor (Fig. 3d), with the exception of parameter 5, all of the parameters have either a negative effect or no significant effect. This complex pattern of sensitivity in temperature and moisture is a reflection of the various processes that impact the surface atmospheric conditions and the difficulties in representing them in the physics parameterizations. The convective parameterization adjusts not only the moisture profiles, but also the temperature profiles. Therefore, it is not surprising to

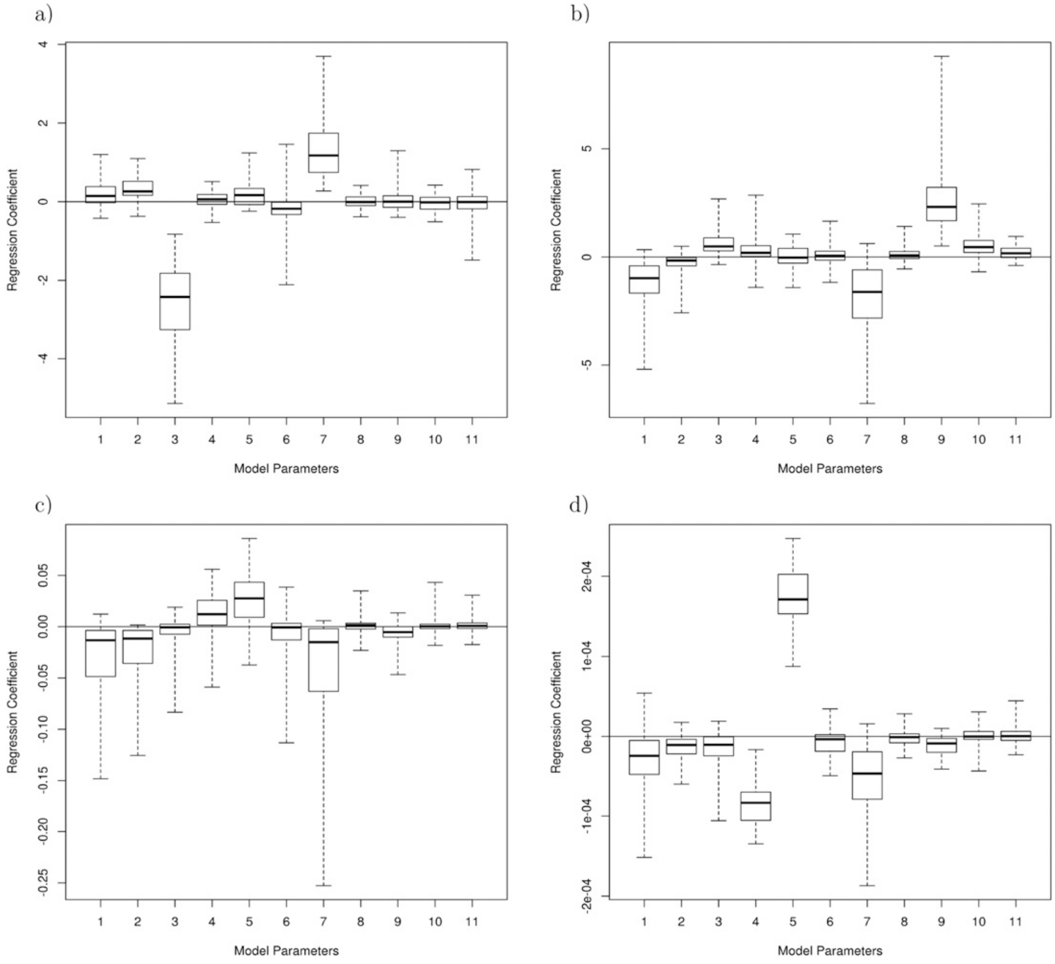

FIG. 3. Sensitivity (on y axis) of the domain mean of (a) convective precipitation, (b) grid-scale precipitation, (c) surface air temperature, and (d) water vapor, with respect to the 11 model parameters (on x axis) listed in Table 1. The boxplots display the variability of the sensitivity across the 40 days examined here.

see that parameters 1, 2, and 7 have impacts on the surface air temperature (Fig. 3c). Parameter 4 modulates the depth of the layer impacted by the vertical mixing processes. For the simulation period, an increase in the vertical mixing length helps to mix heat downward to the surface layer on average, although the impact varies between cooling and heating of the surface layer on particular days. On the other hand, the near-surface water vapor is mixed upward, resulting in a dryer surface layer (Fig. 3d), which is the case for all simulation days. An outstanding feature in Fig. 3d is the high positive sensitivity of water vapor to surface flux (parameter 5). This feature is a strong indication that air moisture in the surface layer is directly influenced by the surface latent flux, which is consistent with the surface physics processes in the model. Likewise, the surface sensible heat flux impacts surface air temperature, indicated by its strong sensitivity to parameter 5 (Fig. 3c).

Precipitation sensitivity analysis was studied using a different (global) method by Marzban et al. (2014), and their results are consistent with those above. This consistency adds support to the results found here and also justifies the use of the simpler (i.e., local or regression based) approach employed here. The aforementioned previous work did not examine the other forecast quantities considered here. These nonspatial sensitivity findings are important because they aid in organizing and discussing the spatial results.

b. Spatial sensitivity

As shown in Fig. 1, one of the outputs of the proposed methodology is a single map of p values that assesses the extent to which any (i.e., at least one) of the model parameters affects the forecasts at any (i.e., at least one) of the grid points in the MMR window. Figure 4 shows such a map for the case where the forecasts are for convective precipitation. A dark grid point corresponds to a small p value, that is, a sensitive grid point. By contrast, a white grid point indicates the lack of evidence from data to support the claim that any of the parameters have an effect on any of the grid points in the MMR window. Clearly, there is nontrivial spatial structure. The grid points in the Southeast and off the eastern coast of the United States appear to be most affected by at least one of the model parameters. There are isolated regions in Mexico and in the northwestern United States that also display sensitivity, albeit to a lesser degree. As explained above, the main utility of such a map is to determine whether or not there exists any effect at all, and whether there is sufficient justification for assessing the contribution of each model parameter, separately. In this case, the existence of an unambiguous spatial structure justifies the latter.

For forecasts of convective precipitation, the maps of the p values for the 11 parameters are shown in Fig. 5.[2] Recall that according to Fig. 3a, parameter 3 (controlling the amount of precipitation fed back to the resolved grid) and parameter 7 (controlling temperature perturbations at the LCL) have the most effect on the domain mean of convective precipitation. From Fig. 5, it is evident that the effect is mostly in the southeastern regions in the forecast domain, although parameter 3 has a larger region of influence than parameter 7. Parameter 2, which also affects KF, has a similar spatial structure. Parameter 5, a surface flux factor, is significant only in the southeastern United States, mainly over the ocean, where large moisture flux occurs across the air–sea interface. The remaining parameters have little or no discernible spatial structure consistent across the 40 days. One possible reason for this independence on surface conditions is that the convective systems included in the data are not strongly surface-driven during the forecast period.

Also, note that the conclusion following from the omnibus test (that at least one of the 11 parameters is responsible for the spatial structure seen in Fig. 4) is consistent with the results in Fig. 5. Indeed, it appears that most of the contribution to the spatial structure in Fig. 4 is from parameter 3.

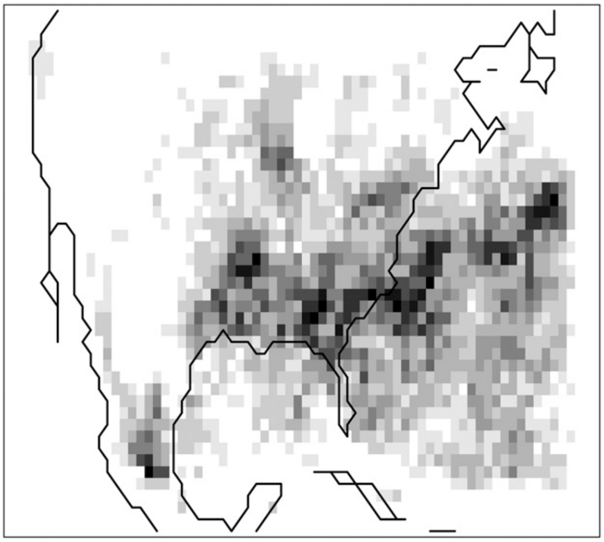

FIG. 4. Sensitivity maps for convective precipitation. Black-colored grid points correspond to small p values, thereby suggesting that at least one of the model parameters has a significant effect on at least one of the grid points in the MMR window. White grid points (large p values) suggest that there is no evidence from the data that any of the model parameters have an effect on any of the grid points in the MMR window. The grayscale colors are assigned to the range of p values across all 11 model parameters.

For grid-scale precipitation, the omnibus map of p values shows an unambiguous spatial structure as well (not shown), and the effect of the individual model parameters is shown in Fig. 6. The "best" parameters, according to Fig. 3b, are parameters 1, 7, and 9, and according to Fig. 6, they all have a comparable spatial structure. In fact, with the exception of parameter 8, which displays no spatial structure, all of the parameters appear to affect the eastern regions of the forecast domain. This spatial structure may be because many of the parameters (e.g., 1 and 7) are KF triggers. They affect the tradeoff between grid-scale precipitation and convective precipitation along the frontal zones, which occur primarily in the eastern to southeastern United States. Parameters 9 (an autoconversion factor) and 10 (slope intercept for rain) are associated with rain production by the microphysics scheme; it is interesting that they are most influential in the northeastern United States, where convective processes are not as active as over Florida. Parameters 4 [planetary boundary layer (PBL) mixing length factor] and 5 (surface flux factor) have similar patterns of influence, with parameter 4 more significant in the northeast and parameter 5 more

[2] It is important to point out that the grayscale colors are assigned to the p values appearing within each panel (i.e., for each model parameter, separately); this color assignment enhances the spatial structure in each panel and is therefore more appropriate for the current study. If the color assignment were made to the p values across all panels (i.e., across all model parameters), then the resulting figures would display little to no information on the spatial structure of the sensitivities. For example, with the alternative color assignment, only the panels corresponding to parameters 3 and 7 in Fig. 5 would display any spatial structure at all because as seen from Fig. 3a, those two parameters dominate the other model parameters. Said differently, the alternative color scheme, like Fig. 3, would be more appropriate for ranking the model parameters, while Fig. 5 is more conducive to an examination of spatial structure of sensitivities.

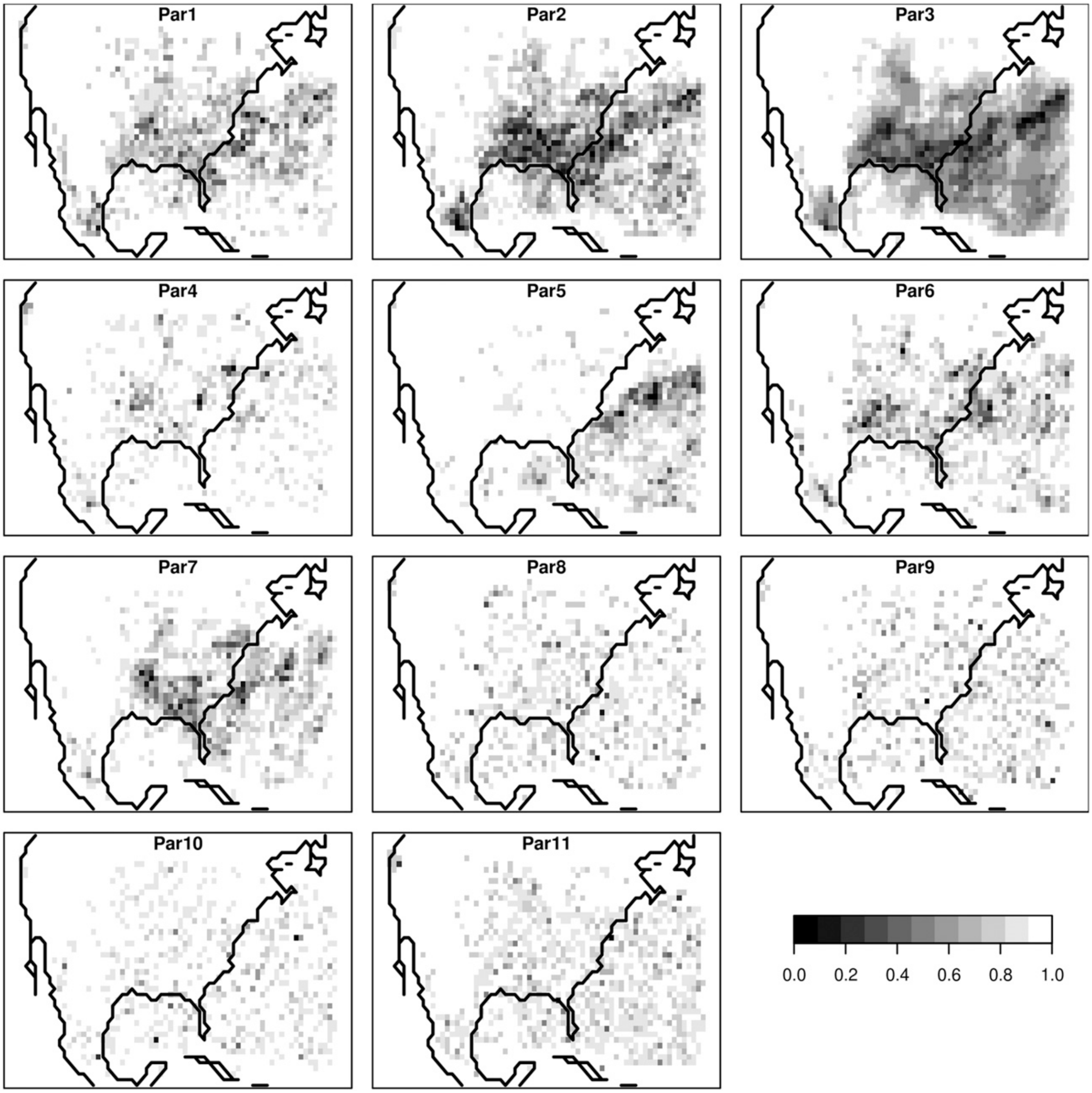

FIG. 5. As in Fig. 4, but for each of the 11 model parameters. Black-colored grid points correspond to small p values, thereby suggesting that the corresponding model parameter has a significant effect on at least one of the grid points in the MMR window.

significant in the east, especially over the Atlantic. This pattern may be an indication of preferred locations for vertical mixing in the boundary layer.

With respect to surface air temperature, Fig. 7 shows the spatial structure of all model parameters, where blue (red) colors correspond to low (high) p values, that is, high (low) sensitivity.[3] Although it is possible to use Fig. 3c as a starting point for discussing Fig. 7, in which case parameters 1, 2, 4, 5, and 7 are the most important,

[3] There are two reasons why the blue/red color scheme is adopted in Fig. 7 (and Fig. 8) instead of the black/white scheme used in Figs. 5 and 6. First, the blue/red color scheme visually enhances the spatial structure. Second, this scheme is more appropriate for continuous fields, such as temperature and water vapor; discrete fields like precipitation have geographic regions where no precipitation may occur at all, to which it is natural to assign the color white, indicating lack of sensitivity. For continuous fields, there exists no geographic region where there exists no field value at all, and so there is no need for a white color.

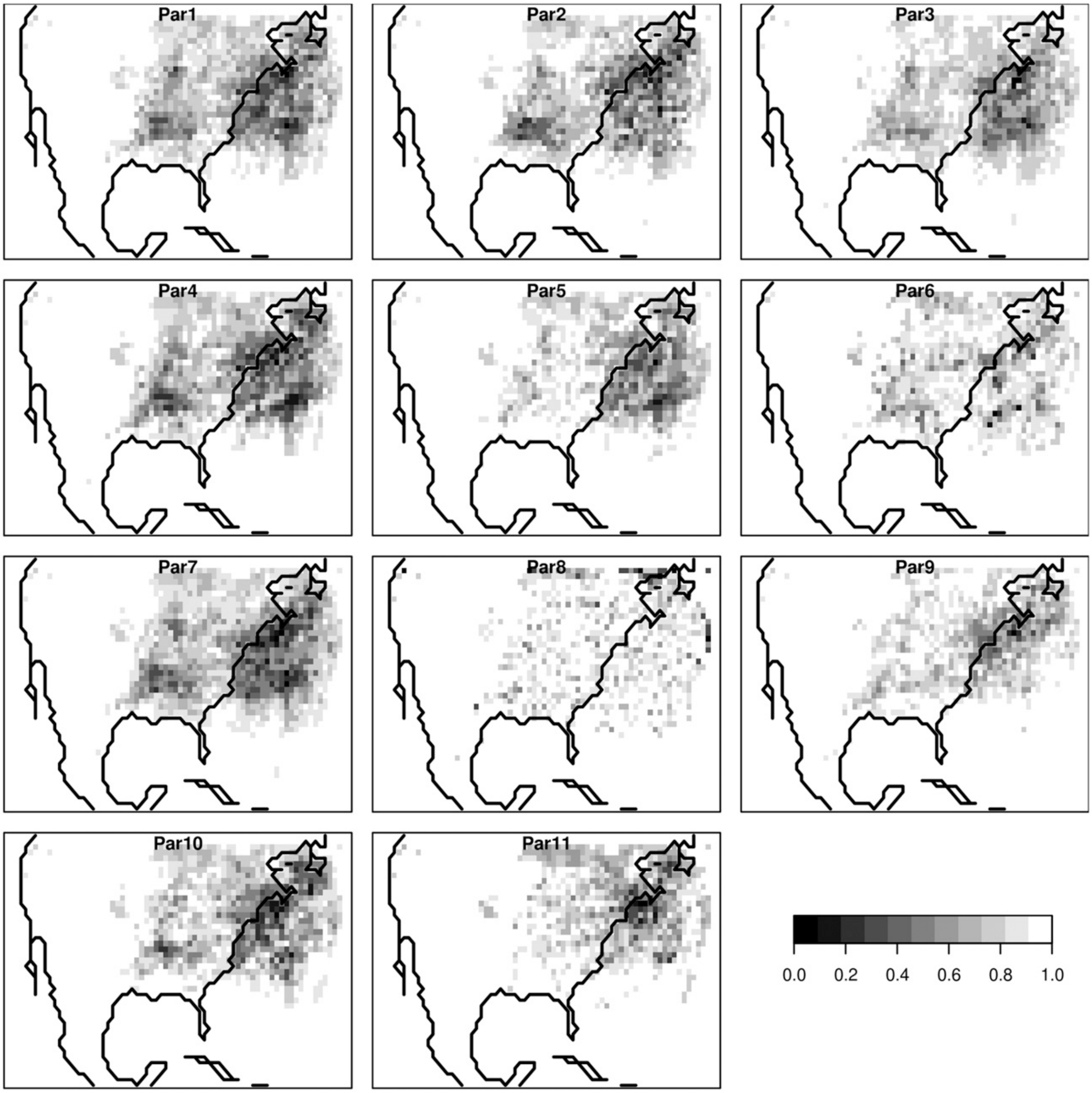

FIG. 6. As in Fig. 5, but for grid-scale precipitation.

that is not necessary because the spatial structures in the various panels in Fig. 7 are sufficiently different to allow for a meaningful discussion independently of the relative strength of the parameters.

Parameters 1, 2, and 7 are related to the KF cumulus parameterization. The largest sensitivity to these parameters is over land, especially over Mexico and over the southeastern United States, where there exists significant convective activity on many of the 40 days examined here. It is interesting that parameter 7 shows a distinct preference for land sensitivity, especially on the entire length of the western coast.

Parameters 4 and 5 are important nearly everywhere across the domain, with the exception of a small pocket of the southern Gulf states. This pattern can be understood by noting that parameters 4 and 5 are the PBL mixing length factor and the surface flux factor, respectively, and so they influence the heat transfer in the boundary layer and across the air–sea/land surface, thus influencing the surface air temperature. The greatest sensitivity to these parameters is in the north and especially over water, where the surface flux and the mixing cause more influence between the sea surface temperature and the air temperature.

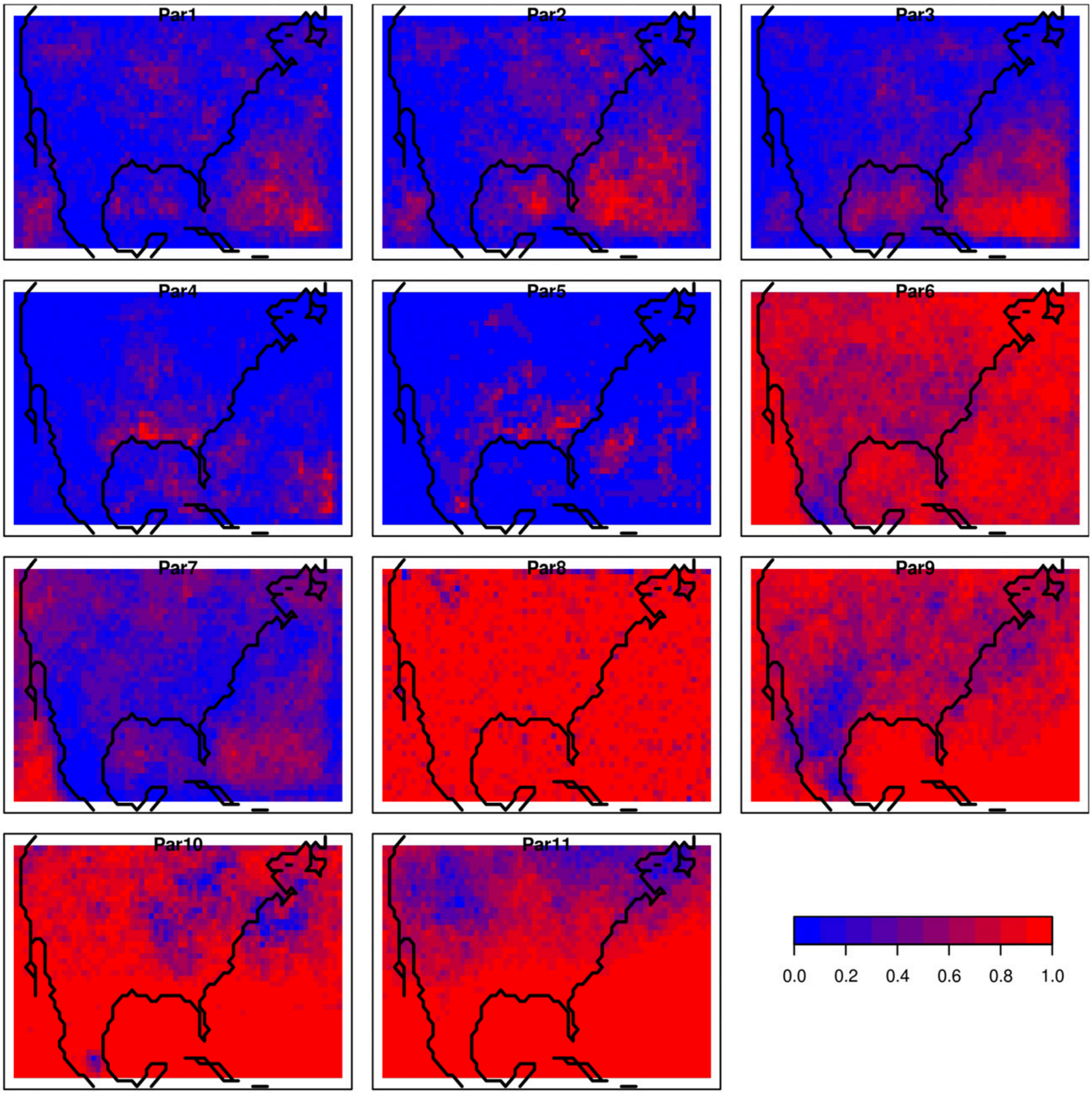

FIG. 7. As in Fig. 5, but for surface air temperature. To enhance visual acuity, a colored scale is used to display the p values. Blue (red) grid points correspond to low (high) p values, that is, a significant (nonsignificant) effect. The colors are assigned to the range of p values across all 11 model parameters.

Parameter 9 (an autoconversion factor for the microphysics) shows sensitivity over the southern Rockies and Mexico. This finding is consistent with the expectation that temperature over that mountainous region is substantially influenced by thermodynamic processes in the clouds, such as condensation heating/evaporation cooling and cloud–radiation interaction.

Parameters 10 and 11 demonstrate a strong spatial pattern of sensitivity in the north. These parameters control size distribution of rain droplets and snow particles, respectively. Over the northeastern United States, physical processes associated with both rain and snow have roughly equal contributions to temperature changes near the surface. However, over the northwestern United States, the snow processes play a dominant role over the rain in modulating the surface air temperature.

The fact that parameters 1–7 all appear to have sensitivity across nearly the entire domain can be attributed to the various physics processes that can directly or indirectly change surface air temperature. For example, evaporative cooling of the rain drops can lower the surface air temperature. The radiative cooling at the cloud top can alter boundary layer structure and result in locations of warming/cooling of air temperature at the surface.

The patterns of spatial sensitivity for forecasts of water vapor (Fig. 8) are similar to those of surface air temperature. Essentially, the sensitivity of water vapor on all parameters, with the exception of parameter 8, appears to have nontrivial spatial structure. This dependence on the model parameters is expected, as all the parameters are related to the availability and conversion of moisture. The only difference between the sensitivity of water vapor and surface air temperature is in the effect of parameters 9–11, where the spatial extent of the sensitive regions is smaller in the former (i.e., the blue regions in Fig. 8 are generally smaller than those in Fig. 7); however, this difference is sufficiently small that it may not be significant.

It is worth emphasizing that parameter 8 (determining the fraction of the clouds to be converted to rain) is not significant for any of the four forecast fields examined, and so it would be reasonable to conclude that it has little to no effect on the spatial structure of forecasts. This lack of sensitivity may be because the grid spacing (81 km) is too coarse to properly resolve the autoconversion process.

4. Summary and discussion

The work reported here describes a proposed methodology for analyzing the spatial structure of the sensitivity of forecasts with respect to model parameters. Although the specific ingredients of the approach are not novel, their application to sensitivity analysis is. The main ingredients are as follows.

(i) Use of Latin hypercube sampling to optimally perturb the model parameters, all assumed to be continuous; the case of discrete or categorical parameters will be reported elsewhere.
(ii) Use of multivariate multiple regression to assess the effect of multiple model parameters on several grid points simultaneously.
(iii) Use of multivariate multiple regression to account for spatial correlations.
(iv) Use of the map of *p* values to display the spatial structure of sensitivities.
(v) Use of the uniformity of *p* values (under the null hypothesis of no effect) to assess the statistical significance of the sensitivities across time (i.e., across 40 days).

Together, these steps lead to a geographic map of *p* values that visually displays the spatial structure of sensitivities for each model parameter.

When applied to the 11 model parameters listed in Table 1, it is found that all of the parameters have significant effects, but the spatial structure of the sensitivities varies with the forecast quantity. The one exception is parameter 8 (an autoconversion factor for the microphysics), which appears to have no effect and no spatial structure at all. The spatial patterns of sensitivity are complex, but can be summarized as follows.

The spatial structure of sensitivities for convective precipitation is most distinct for parameter 3 (fraction of grid-scale precipitation in KF fed back to the grid scale), where the sensitivity is strongest over the southeast of the forecast domain, regardless of whether the forecast is over land or over water. By contrast, the spatial structure of sensitivities for grid-scale precipitation is similar across all of the model parameters (except for parameter 8), and they do show a land–water distinction. Surface air temperature and water vapor have similar spatial structures. Specifically, parameters 1, 2, 3, and 7 (all related to KF) have a strong effect on land but nearly no effect over the oceans. Parameters 9–11 (all related to microphysics) appear to have an influence in three relatively small regions—in the southwest, northeast, and north, respectively. It should be emphasized that because of the coarse model resolution used in the simulation, these sensitivity results (especially those related to precipitation) must be viewed mostly as a demonstration of the methodology developed here.

The spatial structure of sensitivities for grid-scale precipitation (Fig. 5) strongly resembles its climatology (Fig. 2). In other words, it appears that the most sensitive grid points are where the most grid-scale precipitation occurs. First, it is important to emphasize that in spite of this general pattern, the spatial structure of sensitivities does appear to vary across the model parameters. Regardless, the authors have attempted to explain this resemblance by performing a number of different analyses. For example, given that here, sensitivity is measured by regression coefficients, it is possible that the resemblance is due to a mean–variance relationship (Montgomery 2009). Briefly, this condition refers to the situation when the mean of a response is a function of the variance of the errors. When a mean–variance relationship exists in data, not only are the basic probabilistic assumptions of regression violated, but also the variance of the regression coefficients will no longer be constant. In the present application, a mean–variance relationship would lead to a map of *p* values that reflects nothing more than the spatial variability of grid-scale precipitation itself, that is, the climatology. To test the prevalence of a mean–variance relationship, the residual plots were visually examined at each grid point. With the exception of a few (2 to 10, depending on the day being

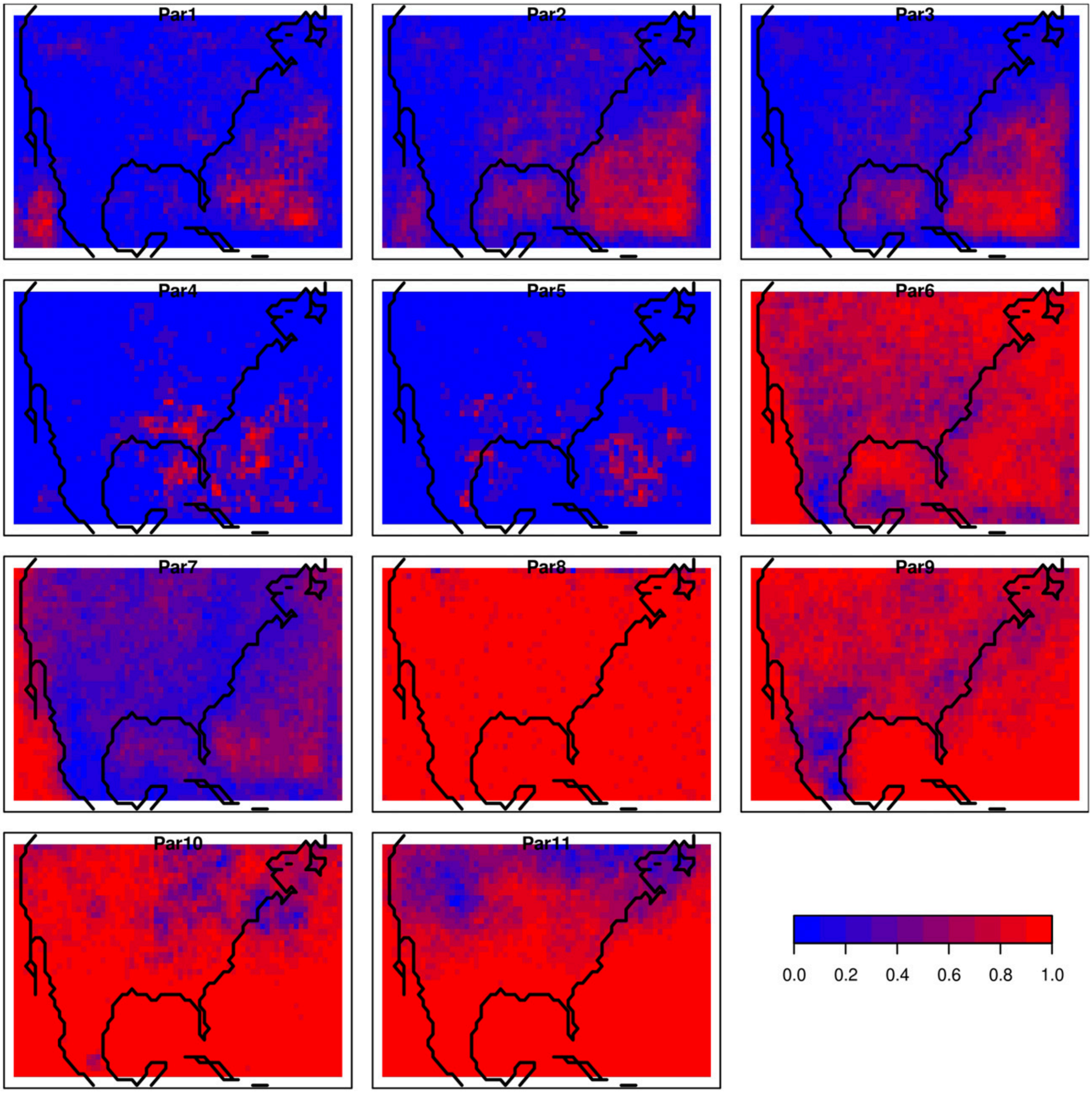

FIG. 8. As in Fig. 7, but for water vapor.

modeled) grid points, no such relationship was noted. And when appropriate data transformations were performed to stabilize the variance, the resulting spatial maps of p values was not affected at all. In short, the similarity of the spatial structure of sensitivities and climatology is not due to a mean–variance relationship. A mean–variance relationship does exist when one examines the scatterplot of mean precipitation versus the variance of precipitation across grid points, as that is a direct consequence of the nearly lognormal distribution of precipitation. However, the mean–variance relationship that is being discussed in the body of the paper refers to the mean and variance of the response across the cases (i.e., the 99 values of the model parameters). It is the latter that is problematic in regression. The former relationship is not problematic because here, no regression is performed across the grid points.

In another attempt to understand the above resemblance, the regression models for sensitivity were also developed with the response set to the anomalies (i.e., forecast − climatology). But, again, the same map of sensitivities was obtained. In summary, the spatial structure of sensitivity in grid-scale precipitation does not appear to be a direct consequence of climatology.

All of the spatial patterns noted in this work deserve some sort of an explanation based on the underlying physics and/or the type of weather events included in the dataset. Although some explanations have been offered, more can be done. For example, it will be useful to cluster the 40 days of data examined here into small groups that are meteorologically homogeneous and repeat the above analysis on each group. Of course, that exercise will lead to generally degraded levels of statistical significance because of the smaller sample sizes in each group; however, that limitation is not a concern because it is the spatial structure of the p values that is under examination, not the magnitude of the p values at each grid point.

On a more technical side, the methodology itself can be extended in several directions. For example, it will be interesting to compare the MMR's ability to account for spatial correlations with that of alternative methods (Douglas et al. 2000; Elmore et al. 2006; Wilks 1997). Although comparisons between the various methods have already been made by DelSole and Yang (2011), such comparisons within the context of sensitivity analysis have not been made.

It may also be important to account for temporal correlations; although the 40 days examined here were taken several days apart in order to minimize temporal dependence, it may be useful to account for any remaining correlation as well. The results are unlikely to change the major conclusions reported here, but they may reveal other spatial structures unseen here. Such hidden spatial structures may also emerge if higher-resolution forecasts are examined, but the results of such an analysis can only reveal smaller-scale structures in the spatial structures already established in this study.

Although all 11 model parameters are used simultaneously in the MMR model, here, the model does not include any interaction terms. There are 55 (i.e., 11 choose 2) such terms, and so a full model with interactions will have 66 (11 + 55) regression coefficients. As such, the size of the Latin hypercube sample may have to be increased beyond 99, and the generation of those data will require more computational effort. To tame the effort, one may introduce interactions only for the parameters that have been found to have an effect on forecasts based on the no-interaction model used here. For example, given that parameter 8 has been found to have generally no effect on any of the forecasts examined here, it is reasonable to assume that it will have little or no interaction with the other parameters either. This assumption is generally borne out due to several principles: the principle of hierarchical ordering, the principle of effect sparsity, and the principle of effect hierarchy (Montgomery 2009, p. 192, 230, 272, 314, 329; Li et al. 2006, 33–34). Alternatively, one may choose to include only interactions that affect only the domain mean of the forecast quantities; for mean convective precipitation, all interactions are already analyzed in Marzban et al. (2014).

In all methods that rely on a "window" of some kind, the window size usually requires some consideration. Here, the MMR window has a fixed size (3 × 3), and that size is selected by qualitative considerations; for example, it is large enough to account for some spatial structure and small enough to yield a map with a reasonably large number of grid points. However, it is possible to allow for the window size to vary across the spatial domain, adaptively, with the size determined by an estimate of the spatial correlation (e.g., via variograms; see Cressie 1993).

It is worth pointing out that the role played by the MMR window is more than simply smoothing the map of p values. Although the very existence of an MMR window does lead to smoother maps, that smoothing is only a side effect. The main purpose of the MMR window is to account for spatial correlations in assessing sensitivity. As mentioned above, failure to take such correlations into account leads to wrong (biased) estimates of sensitivity. Another way to highlight the effect of the MMR window is to note that a generic smoother would operate on the p values directly; by contrast, any smoothing that may be occurring in MMR is based on the correlations across grid points and between model parameters.

Acknowledgments. This work has received support from the Office of Naval Research (N00014-12-G-0078 Task 29) and National Science Foundation (AGS-1402895). We are grateful to Paul Roebber for a discussion on the differences between a sensitivity analysis and model tuning. Ethan P. Marzban is acknowledged for generating the flowchart in Fig. 1.

REFERENCES

Adlerman, E. J., and K. K. Droegemeier, 2002: The sensitivity of numerically simulated cyclic mesocyclogenesis to variations in model physical and computational parameters. *Mon. Wea. Rev.*, **130**, 2671–2691, https://doi.org/10.1175/1520-0493(2002)130<2671:TSONSC>2.0.CO;2.

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.

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.

Boyle, J. S., S. A. Klein, D. D. Lucas, H.-Y. Ma, J. Tannahill, and S. Xie, 2015: The parametric sensitivity of CAM5's MJO. *J. Geophys. Res. Atmos.*, **120**, 1424–1444, https://doi.org/10.1002/2014JD022507.

Bretz, F., T. Hothorn, and P. Westfall, 2010: *Multiple Comparisons Using R*. Chapman and Hall/CRC, 208 pp.

Cacuci, D. G., M. Ionescu-Bujor, and I. M. Navon, 2005: *Sensitivity and Uncertainty Analysis, Volume II: Applications to Large-Scale Systems*. Chapman and Hall/CRC, 368 pp.

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.

Crook, N. A., 1996: Sensitivity of moist convection forced by boundary layer processes to low-level thermodynamic fields. *Mon. Wea. Rev.*, **124**, 1767–1785, https://doi.org/10.1175/1520-0493(1996)124<1767:SOMCFB>2.0.CO;2.

Daescu, D. N., and R. H. Langland, 2013: Error covariance sensitivity and impact estimation with adjoint 4D-Var: Theoretical aspects and first applications to NAVDAS-AR. *Quart. J. Roy. Meteor. Soc.*, **139**, 226–241, https://doi.org/10.1002/qj.1943.

Davis, C. A., and K. Emanuel, 1991: Potential vorticity diagnostics of cyclogenesis. *Mon. Wea. Rev.*, **119**, 1929–1953, https://doi.org/10.1175/1520-0493(1991)119<1929:PVDOC>2.0.CO;2.

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

Dmitrienko, A., F. Bretz, P. H. Westfall, J. Troendle, B. L. Wiens, A. C. Tamhane, and J. C. Hsu, 2009: Multiple testing methodology. *Multiple Testing Problems in Pharmaceutical Statistics*, A. Dmitrienko, A. Tamhane, and F. Bretz, Eds., Chapman and Hall, 35–98.

Douglas, E. M., R. M. Vogel, and C. N. Kroll, 2000: Trends in floods and low flows in the United States: Impact of spatial correlation. *J. Hydrol.*, **240**, 90–105, https://doi.org/10.1016/S0022-1694(00)00336-X.

Duan, Q., and Coauthors, 2006: Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops. *J. Hydrol.*, **320**, 3–17, https://doi.org/10.1016/j.jhydrol.2005.07.031.

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

Fang, K.-T., R. Li, and A. Sudjianto, 2006: *Design and Modeling for Computer Experiments*. Chapman & Hall/CRC, 290 pp.

Fox, J., M. Friendly, and S. Weisberg, 2013: Hypothesis tests for multivariate linear models using the car package. *The R J.*, **5**, 39–52.

Gallus, W. A., Jr., 1999: Eta simulations of three extreme precipitation events: Sensitivity to resolution and convective parameterization. *Wea. Forecasting*, **14**, 405–426, https://doi.org/10.1175/1520-0434(1999)014<0405:ESOTEP>2.0.CO;2.

Gombos, D., and J. A. Hansen, 2008: Potential vorticity regression and its relationship to dynamical piecewise inversion. *Mon. Wea. Rev.*, **136**, 2668–2682, https://doi.org/10.1175/2007MWR2165.1.

Gómez-Navarro, J. J., C. C. Raible, and S. Dierer, 2015: Sensitivity of the WRF Model to PBL parametrisations and nesting techniques: Evaluation of wind storms over complex terrain. *Geosci. Model Dev.*, **8**, 3349–3363, https://doi.org/10.5194/gmd-8-3349-2015.

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*, **63**, 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.

Holt, T. R., J. A. Cummings, C. H. Bishop, J. D. Doyle, X. Hong, S. Chen, and Y. Jin, 2011: Development and testing of a coupled ocean–atmosphere mesoscale ensemble prediction system. *Ocean Dyn.*, **61**, 1937–1954, https://doi.org/10.1007/s10236-011-0449-9.

Hourdin, F., and Coauthors, 2017: The art and science of climate model tuning. *Bull. Amer. Meteor. Soc.*, **98**, 589–602, https://doi.org/10.1175/BAMS-D-15-00135.1.

Houston, A. L., and D. Niyogi, 2007: The sensitivity of convective initiation to the lapse rate of the active cloud-bearing layer. *Mon. Wea. Rev.*, **135**, 3013–3032, https://doi.org/10.1175/MWR3449.1.

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.

Kain, J. S., and J. M. Fritsch, 1993: Convective parameterization for mesoscale models: The Kain–Fritsch scheme. *The Representation of Cumulus Convection in Numerical Models, Meteor. Monogr.*, No. 46, Amer. Meteor. Soc., 165–170.

Kucherenko, S., D. Albrecht, and A. Saltelli, 2015: Exploring multi-dimensional spaces: A comparison of Latin hypercube and quasi Monte Carlo sampling techniques. Cornell University Rep., 30 pp., https://arxiv.org/abs/1505.02350.

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.

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.

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.

——, S. Sandgathe, 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.

McKay, M. D., R. J. Beckman, and W. J. Conover, 1979: A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, **21**, 239–245, https://doi.org/10.2307/1268522.

Mölders, N., M. Laube, and G. Kramm, 1995: On the parameterization of ice microphysics in a mesoscale α weather forecast model. *Atmos. Res.*, **38**, 207–235, https://doi.org/10.1016/0169-8095(94)00094-T.

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

Noble, W. S., 2009: How does multiple testing correction work? *Nat. Biotechnol.*, **27**, 1135–1137, https://doi.org/10.1038/nbt1209-1135.

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.

Qian, Y., and Coauthors, 2015: Parametric sensitivity analysis of precipitation at global and local scales in the Community Atmosphere Model CAM5. *J. Adv. Model. Earth Syst.*, **7**, 382–411, https://doi.org/10.1002/2014MS000354.

Robock, A., and Coauthors, 2003: Evaluation of the North American Land Data Assimilation System over the southern Great Plains during warm seasons. *J. Geophys. Res.*, **108**, 8846, https://doi.org/10.1029/2002JD003245.

Roebber, P. J., 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.

——, and L. F. 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.

Rosenblatt, J., 2013: A practitioner's guide to multiple hypothesis testing error rates. Cornell University Rep., 32 pp., https://arxiv.org/abs/1304.4920.

Rutledge, S. A., and P. Hobbs, 1983: The mesoscale and microscale structure and organization of clouds and precipitation in midlatitude cyclones. VIII: A model for the "seeder-feeder" process in warm-frontal rainbands. *J. Atmos. Sci.*, **40**, 1185–1206, https://doi.org/10.1175/1520-0469(1983)040<1185:TMAMSA>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.

Saltelli, A., S. Tarantola, F. Campolongo, and M. Ratto, 2004: *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*. John Wiley & Sons, 232 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. I. Notz, 2003: *The Design and Analysis of Computer Experiments*. Springer, 283 pp.

Schumann, M. R., and P. J. Roebber, 2010: The influence of upper-tropospheric potential vorticity on convective morphology. *Mon. Wea. Rev.*, **138**, 463–474, https://doi.org/10.1175/2009MWR3091.1.

Stensrud, D. J., 2007: *Parameterization Schemes: Keys to Understanding Numerical Weather Models.* Cambridge University Press, 459 pp.

Torn, R. D., and G. Hakim, 2008: Ensemble-based sensitivity analysis. *Mon. Wea. Rev.*, **136**, 663–677, https://doi.org/10.1175/2007MWR2132.1.

Voudouri, A., P. Khain, I. Carmona, O. Bellprat, F. Grazzini, E. Avgoustoglou, and J. M. Bettems, and P. Kaufmann, 2017: Objective calibration of numerical weather prediction models. *Atmos. Res.*, **190**, 128–140, https://doi.org/10.1016/j.atmosres.2017.02.007.

Warner, T. T., 2011: *Numerical Weather and Climate Prediction*. Cambridge University Press, 526 pp.

Weisman, M. L., and Coauthors, 2015: The Mesoscale Predictability Experiment (MPEX). *Bull. Amer. Meteor. Soc.*, **96**, 2127–2149, https://doi.org/10.1175/BAMS-D-13-00281.1.

Wilks, D. S., 1997: Resampling hypothesis tests for autocorrelated fields. *J. Climate*, **10**, 65–82, https://doi.org/10.1175/1520-0442(1997)010<0065:RHTFAF>2.0.CO;2.

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

——, 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.

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.