# Variance-Based Sensitivity Analysis: Preliminary Results in COAMPS

CAREN MARZBAN

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

SCOTT SANDGATHE

*Applied Physics Laboratory, University of Washington, Seattle, Washington*

JAMES D. DOYLE

*Naval Research Laboratory, Monterey, California*

NICHOLAS C. LEDERER

*The Boeing Company, Applied Mathematics, Seattle, Washington*

ABSTRACT

Numerical weather prediction models have a number of parameters whose values are either estimated from empirical data or theoretical calculations. These values are usually then optimized according to some criterion (e.g., minimizing a cost function) in order to obtain superior prediction. To that end, it is useful to know which parameters have an effect on a given forecast quantity, and which do not. Here the authors demonstrate a variance-based sensitivity analysis involving 11 parameters in the Coupled Ocean–Atmosphere Mesoscale Prediction System (COAMPS). Several forecast quantities are examined: 24-h accumulated 1) convective precipitation, 2) stable precipitation, 3) total precipitation, and 4) snow. The analysis is based on 36 days of 24-h forecasts between 1 January and 4 July 2009. Regarding convective precipitation, not surprisingly, the most influential parameter is found to be the fraction of available precipitation in the Kain–Fritsch cumulus parameterization fed back to the grid scale. Stable and total precipitation are most affected by a linear factor that multiplies the surface fluxes; and the parameter that most affects accumulated snow is the microphysics slope intercept parameter for snow. Furthermore, all of the interactions between the parameters are found to be either exceedingly small, or have too much variability (across days and/or parameter values) to be of primary concern.

## 1. Introduction

Sensitivity analysis (SA) refers to a wide suite of techniques for assessing the effect of a set of quantities on another. In meteorological circles, the latter (here called *output*) is usually some forecast quantity of interest (e.g., total accumulated precipitation, 2-m temperature, 10-m wind speed, etc.). The former (here called *input*), is usually either analyzed initial conditions or model/algorithm parameters, or both. In some cases the input is an observation, an entire class of observations, or the specification of background errors in data assimilation, whose effect is of interest. Approaches based on adjoints, and ensembles, or both have been proposed for performing SA (Ancell and Hakim 2007; Daescu and Langland 2013; Davis and Emanuel 1991; Gombos and Hansen 2008; Hacker et al. 2011; Torn and Hakim 2008).

The approaches tend to fall into two broad (not mutually exclusive) categories; some are designed to deal with a large number of inputs in a numerically efficient manner. Adjoint-based models are one such example (Ancell and Hakim 2007). Indeed, a data assimilation system can be said to perform SA, albeit only implicitly. Alternative approaches are better suited to smaller number of inputs, and may even be more computationally intensive. Variance-based SA methods belong to this category (McKay et al. 1979; Oakley and O'Hagan 2004; Saltelli et al. 2008, 2010; Santner et al. 2003; Sobol 1993). The value of these approaches is in providing a

_____

*Corresponding author address:* Caren Marzban, Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322.
E-mail: marzban@stat.washington.edu

more in-depth assessment of the relationship between the input and the output. The results can have practical implications (e.g., identifying the parameters whose values must be set carefully and those deserving of less attention). Knowledge of both sets of parameters is important; setting the value of an influential parameter deserves care and attention, while far less attention can be paid to the task of setting the value of a noninfluential parameter. Variance-based methods can also lead to a better theoretical understanding of the nature of the underlying relationship between the input and output (Bolado-Lavin and Badea 2008; Marzban 2011, 2013).

In a recent effort (Holt et al. 2011) 11 parameters in the Coupled Ocean–Atmosphere Mesoscale Prediction System (COAMPS)[1] were varied for the purpose of studying their effect on the forecasts.[2] Here, a variance-based method is applied to that problem. Variance-based methods are ideally suited to *computer experiments* (e.g., data generated from a computer model, resulting from changing the model parameters; Bowman et al. 1993; Fang et al. 2006; Sacks et al. 1989a,b; Welch et al. 1992).[3] Here, the choice of the parameters is the same as that of Holt et al. (2011), but the manner in which their values are varied is different. Whereas in Holt et al. a handful of different values were selected about the mean of each parameter, here a method of sampling is employed which is known to have desirable properties; for example, it is known to provide more precise (at least, no less precise) estimates than simple random sampling. A variance-based SA method, along with this sampling scheme, was recently demonstrated on the Lorenz'63 model (Marzban 2011, 2013). The main goals of this paper are 1) to demonstrate the application of the variance-based SA methodology to an operational weather forecasting model and 2) to study the effect of the 11 parameters on several types of precipitation and snow forecasts.

## 2. Data

Only the atmospheric portion of COAMPS (Hodur 1997) is used in this study. The COAMPS model is based on a finite-difference approximation to the fully compressible, nonhydrostatic equations and uses a terrain-following vertical coordinate transformation. The compressible equations are integrated using a time-splitting technique with a semi-implicit formulation for the vertical acoustic modes (Klemp and Wilhelmson 1978).

A prognostic equation for the turbulence kinetic energy budget is used to represent the planetary boundary layer and free-atmospheric turbulent mixing and diffusion (Hodur 1997). The Louis (1979) surface-layer parameterization, which makes use of a surface energy budget based on the force-restore method, is used to represent the surface fluxes. Subgrid-scale moist convection is represented using the Kain and Fritsch (1993) parameterization. The grid-scale evolution of the moist processes is predicted explicitly from budget equations for cloud water, cloud ice, raindrops, snowflakes, and water vapor following Rutledge and Hobbs (1983). The shortwave and longwave radiation processes are represented following Harshvardhan et al. (1987). Results from COAMPS model simulations have been evaluated on numerous occasions using special observations and field campaign datasets and have been demonstrated to accurately simulate mesoscale flows (e.g., Hodur 1997; Doyle et al. 2011; Jiang and Doyle 2009).

The forecast data used in this study are produced using COAMPS version 4.2.2 run at the Applied Physics Laboratory, University of Washington. The COAMPS model was forced using 0.5° resolution initial and one-way boundary conditions from the Navy Operational Global Atmospheric Prediction System (NOGAPS). The COAMPS analysis domain is shown in Fig. 1 (white frame) and is based on NOGAPS initial fields and local observations. For computational efficiency COAMPS was run with a grid spacing of 81 km.

The model parameters examined here are shown in Table 1. Computer data are generated by sampling 99 points from the 11-dimensional space of the parameters (called the empirical region).[4] The sampling is done via the Latin hypercube sampling (LHS) method described in Marzban (2011, 2013). Briefly, this type of sampling is designed to assure that no two of the 99 points have the same value for any of the 11 parameters. In other words, the sampling scheme is designed to avoid any clustering of the 99 points, an outcome which is not ruled out under random sampling. The result of such sampling is more precise estimates (at least, no less precise estimates) than random sampling. For this and other reasons, LHS is usually the method of choice in SA of computer experiments (Cioppa and Lucas 2007; Douglas 2005;

---

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

[2] Holt et al. explored additional parameters, particularly in the planetary boundary layer and microphysics parameterizations. The parameters were identified through sensitivity tests and consultation with the parameterization developers. The 11 parameters were found to be the most important parameters in generating ensemble spread and producing reasonable spread–skill relationships.

[3] Data generated from computer experiments are characterized by the absence of variability across repeated runs. They are distinct from "physical experiments" where repeating an experiment generally leads to different results.

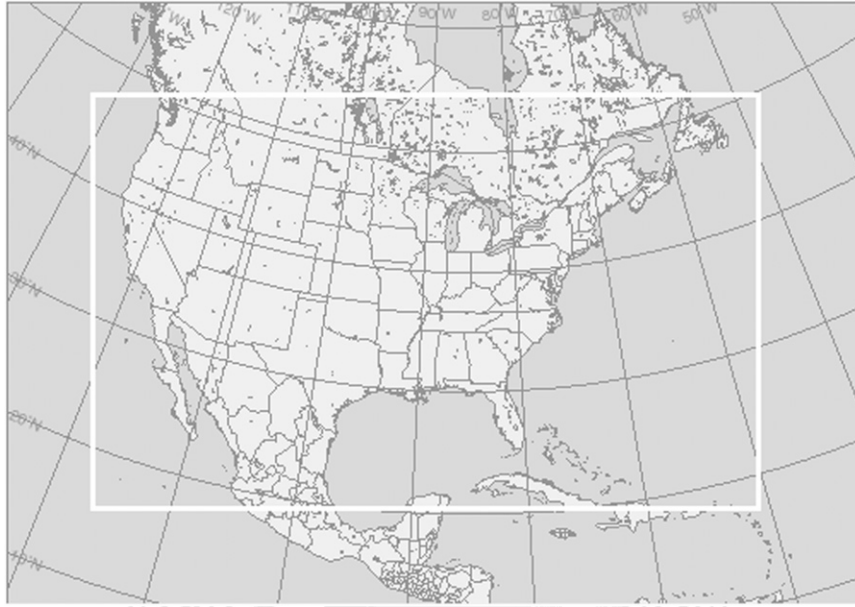[4] Sampling 50 points resulted in similar conclusions.

FIG. 1. The spatial domain (white rectangle) of the analysis.

Marzban 2011, 2013). As in Marzban (2011, 2013), LHS is also compared with simple random sampling (SRS).

For each of 36 dates, beginning with 1 January and ending with 4 July 2009, at approximately 4-day intervals, a different set of 99 points are selected, and COAMPS is run at each of the 99 points in the empirical region.[5] The specific dates are given in Table 2. The 4-day interval is selected to enhance the independence of the data from day to day. For each run, different types of 24-h accumulated precipitation are considered: convective, grid scale (or stable), total (convective plus stable), and total accumulated snow. These quantities are therefore the aforementioned forecast quantities (or outputs), whose dependence on the model parameters (i.e., inputs) is the subject of this work. These four types of precipitation will be denoted by the symbols conv, stab, total, and snow, respectively. These forecast quantities are selected because the data analyzed here begins on 1 January (i.e., when there is significant snow over some parts of the region), and so it is possible that different values of the parameters will shift snow to rain or vice versa. Only "heavy" precipitation is examined—specifically, the 90th percentile across the spatial domain.

## 3. Method: Variance-based SA

A thorough assessment of variance-based SA, as well as a comparison with other SA methods, has been

performed by Bolado-Lavin and Badea (2008). Suffice it to say that the main distinguishing characteristic of these methods is that they rely on probability distributions to describe the inputs and the outputs. As a result, one disadvantage is that they tend to be computationally intensive compared to methods that are not based on probability distributions. However, that disadvantage is compensated by the ability of the method to produce estimates of uncertainty as well.

Let the output (i.e., the forecast quantity) be denoted by $y$ and the inputs (i.e., model parameters) be denoted by $x_1, x_2, \ldots, x_n$. To statistically quantify the mathematical function $y = \eta(x_1, x_2, \ldots, x_n)$, one employs the high-dimensional model representation (Sobol 1993), where the function is written in terms of expected values:

$$y = \eta(x_1, x_2, \ldots, x_n) = E(y) + \sum_{i}^{n} z_i(x_i)$$
$$+ \sum_{i<j} z_{ij}(x_i, x_j) + \cdots, \tag{1}$$

where

$$z_i(x_i) = E(y \,|\, x_i) - E(y), \tag{2}$$

$$z_{ij}(x_i, x_j) = E(y \,|\, x_i, x_j) - E(y \,|\, x_i) - E(y \,|\, x_j) + E(y). \tag{3}$$

The $z_i(x_i)$ is referred to as the *main effect*, and the $z_{ij}(x_i, x_j)$ is called the first-order *interaction* between the

---

[5] Consequently, across the 36 days, the number of points sampled in the empirical region is in fact much larger (i.e., $36 \times 99 = 3564$).

TABLE 1. The 11 parameters studied in this paper. Also shown are the default values and the range over which they are varied. Kain–Fritsch (KF), planetary boundary layer (PBL), and lifting condensation level (LCL).

| 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 ($kg\,m^{-3}\,s^{-1}$) | Autoconversion factors for the microphysics | 0.001 | $1 \times 10^{-4}, 1 \times 10^{-2}$ |
| 9 | autocon2 ($kg\,m^{-3}\,s^{-1}$) | Autoconversion factors for the microphysics | $4 \times 10^{-4}$ | $4 \times 10^{-5}, 4 \times 10^{-3}$ |
| 10 | rainsi ($m^{-1}$) | Microphysics slope intercept parameter for rain | $8.0 \times 10^6$ | $8.0 \times 10^5, 8.0 \times 10^7$ |
| 11 | snowsi ($m^{-1}$) | Microphsyics slope intercept parameter for snow | $2.0 \times 10^7$ | $2.0 \times 10^6, 2.0 \times 10^8$ |

$i$th and $j$th input (parameter). Higher-order terms are assumed to be negligible.

All of the quantities in (1)–(3) are written in terms of $E(y\,|\,x_i)$ and $E(y\,|\,x_i, x_j)$, and so these conditional expected values must be estimated from data on $y$ and $x_i$. In this article, for each of the 36 days, the necessary data are generated by the LHS and SRS procedures mentioned in the previous section. It is known that an ordinary least squares fit to data on $y$ (as response) and $x_i$ (as predictor) provides an estimate of $E(y\,|\,x_i)$ (Draper and Smith 1998). Similarly, a least squares fit to $y$ and two predictors $x_i$ and $x_j$ yields an estimate for $E(y\,|\,x_i, x_j)$. Although more sophisticated methods exist for estimating these expected values (see the summary and discussion section), this regression approach is the one used in the current work. Regardless of how the expected values are estimated, an issue deserving of attention is the complexity of the underlying statistical models; a very complex model, with nonlinear terms and interaction terms, can lead to overfitting, while an overly simple model may not be able to capture possible nonlinearities in the data. Here, a polynomial regression of order two is used for estimating all expected values. Specifically, to estimate $E(y\,|\,x_i)$, the regression model $y = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2$ is used and to estimate $E(y\,|\,x_i, x_j)$, the model is $y = \alpha_0 + \alpha_1 x_i + \beta_1 x_j + \alpha_2 x_i^2 + \beta_2 x_j^2 + \gamma x_i x_j$. Even the more complex of these two models has only 6 coefficients, and so, 99 cases (i.e., size of the empirical region) are adequate for their estimation. As such, the models have some nonlinearity, but not so much that would readily allow overfitting.

A number of sensitivity measures have been proposed in the literature all based on a decomposition of functions called the high-dimensional model representation

(Saltelli et al. 2010; Sobol 1993; Marzban 2011, 2013).[6] Here for simplicity, focus is placed on only two:

$$S_i = \frac{V[E(y\,|\,x_i)]}{V[y]}, \quad S_{ij} = \frac{V[z_{ij}(x_i, x_j)]}{V[y]}. \quad (4) \quad \boxed{\text{AU2}}$$

The measure $S_i$ gauges the expected reduction in the variance of $y$, given $x_i$, expressed as the proportion of the total variance in $y$, $V[y]$. Here $S_{ij}$ measures the interaction between $x_i$ and $x_j$, again expressed as a proportion. Under ideal circumstances when the $x_i$ are orthogonal and the higher-order terms in (1) can be neglected, then $S_i$ and $S_{ij}$ completely decompose the total variance of $y$. As such, they offer a reasonably complete assessment of the sensitivity of all the parameters and their interactions. The quantities $S_i$ and $S_{ij}$ are essentially normalized versions of $z_i$ and $z_{ij}$, and are therefore referred to as main effects and interactions, respectively.

Given the above-mentioned design of the experiment, the sensitivity measures are subject to two sources of variability: daily variability and sampling variability in parameter space. To estimate the former, all sensitivity measures are computed for each of the 36 days in the dataset. To produce a single value of a sensitivity measure, 99 points in the 11-dimensional parameter space are selected. Therefore, an estimate of the variability due to sampling in parameter space is provided by taking multiple (here 20) sets of the 99 parameter values. Given that the main aim of this work is to identify

---

[6] A special case of such a decomposition for binary variables has been developed by Stein and Alpert (1993).

TABLE 2. The dates of the 36 days on which the SA is performed.

| 1 Jan 2009 | 5 Jan 2009 | 17 Jan 2009 | 21 Jan 2009 | 25 Jan 2009 | 29 Jan 2009 |
|---|---|---|---|---|---|
| 6 Feb 2009 | 10 Feb 2009 | 14 Feb 2009 | 18 Feb 2009 | 22 Feb 2009 | 26 Feb 2009 |
| 2 Mar 2009 | 6 Mar 2009 | 10 Mar 2009 | 14 Mar 2009 | 18 Mar 2009 | 22 Mar 2009 |
| 26 Mar 2009 | 30 Mar 2009 | 3 Apr 2009 | 7 Apr 2009 | 11 Apr 2009 | 19 Apr 2009 |
| 23 Apr 2009 | 27 Apr 2009 | 1 May 2009 | 9 May 2009 | 13 May 2009 | 17 May 2009 |
| 21 May 2009 | 25 May 2009 | 29 May 2009 | 22 Jun 2009 | 30 Jun 2009 | 4 Jul 2009 |

important (and unimportant) parameters across all days and all parameter values, the two sources of variability are combined, and the resulting histograms are summarized and presented as boxplots. A comparison of the two sources of variability is also presented.

The assessment of the "statistical significance" is not straightforward when dealing with computer experiments, because the very notion of statistical significance presumes the existence of an experimental error, which in computer experiments does not exist.[7] In other words, in the current data $V[y \mid x_1, x_2, \ldots, x_{11}] = 0$.[8] No attempt is made to present a general procedure for hypothesis testing. The only question that is addressed is whether the variability in the sensitivity measure for a given model parameter is sufficiently large to preclude a reliable conclusion regarding the sensitivity of that parameter.

To be more specific, consider the values of $S_{ij}$ one may obtain for 1 day. For 11 parameters, there will be $(11 \times 10)/2 = 55$ such values. By chance alone, some of these values will be large (near 1), while others will be small (near 0). The question is whether the variability of each of the 55 $S_{ij}$ values, for fixed $i$, $j$, is sufficiently small to suggest that the interaction exists (across all days and all parameter values). To answer that question, the histogram of *all* values of $S_{ij}$ (across 36 days and across 55 pairwise comparisons) is used to approximate its sampling distribution under the null hypothesis of no interaction. This approximation will be accurate if the true number of interacting parameters is small. As will be shown below, the observed pattern of the distributions of $S_{ij}$ suggests that the approximation is reasonable.

Oakley and O'Hagan (2004) propose that it is useful to display the functional form of the dependence of the forecast quantity on the parameters. For that purpose, the $E(y \mid x_i)$ are plotted versus each of the 11 parameters (standardized to have zero mean and one standard deviation). Viewing $E(y \mid x_i, x_j)$ is difficult because of its three-dimensional nature, and so will not be considered here.

## 4. Demonstration: Two parameters

Before embarking on the analysis of the 11-parameter sensitivity analysis, it is helpful to examine a 2-parameter situation. This allows for the visualization of the problem in three dimensions. Also, and again for visual acuity, only 50 points (instead of 99) are selected from the empirical region. For example, Fig. 2 shows the result when the empirical region is spanned only by the fraction of precipitation fed back to the grid scale in the Kain–Fritsch parameterization (prcpfrac) and a linear factor that changes the PBL mixing length (mixlen) (Table 1), and when the forecast quantity is the 90th percentile (across the domain) of 24-h accumulated convective precipitation (conv). The 50 larger circles display the data generated according to the LHS method described previously, only for one day (1 January 2009). The fact that these data are generated from a computer experiment is evident in the manner in which they all reside on a relatively simple surface—in this case a plane. Had there been any experimental error, these points would have fallen not exactly on the surface, but scattered about it.[9] Note that, according to this figure, convective precipitation is sensitive to both prcpfrac and mixlen (as evidenced by the nonzero slopes in both directions), but much more so on the former. This result makes intuitive sense since the changes to the feedback to the grid scale most directly impact the convective precipitation, more so than the boundary layer mixing length.

Different insight is gained from viewing the data as two-dimensional scatterplots. Figure 3 shows the result for 20 days (with different colors); only 20 of the 36 days are shown so as to avoid cluttering the graphs. From the

---

[7] Of course, the COAMPS predictions are subject to initial analysis error, model error, etc. However, those errors are less of a concern because focus is placed only on the manner in which the predictions vary with the parameters.

[8] The variability that is minimized in the aforementioned regression models is due to the variability in $y$ brought about by the changes in the parameters *not* used in the regression models.

[9] Technically, data from a computer experiment are not required to reside on a smooth surface. In other words, it is entirely possible that the squares in Fig. 2 would have not fallen on a plane; in that case, one would require a more complex model for estimating $E(y \mid x_1, x_2)$.
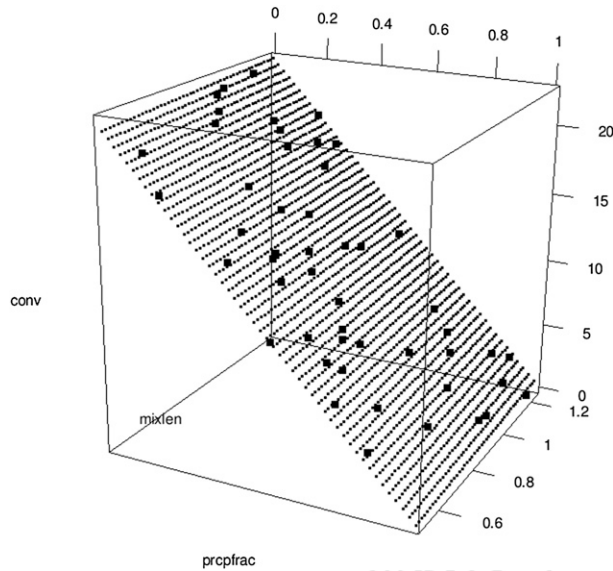
FIG. 2. The response surface of convective precipitation (conv) vs two model parameters: prcpfrac and mixlen. The 50 large squares are the data generated from the computer experiment, all residing on a plane (displayed as small points).

top-left panel, it can be seen that the above-mentioned strong dependence of convective precipitation on prcpfrac exists across all days. As for the above-noted weaker dependence on mixlen, here it is reflected in the larger scatter; that dependence is also persistent across days, as evidenced by the manner in which the various-colored curves are correlated.

The remaining panels in Fig. 3 show analogous results but for stable (top right), total precipitation (bottom left), and accumulated snow (bottom right). In contrast to the relationship between convective precipitation and prcpfrac, where the former decreases with increasing prcpfrac, the effect of prcpfrac on stable precipitation is reversed, and not as pronounced. Once again these results are expected: increasing the feedback from the convective to the grid-scale precipitation should decrease the convective precipitation and increase the grid-scale precipitation. The effect of these two parameters on total precipitation is a combination of the effects on convective and stable precipitation: the linear trend dependence of total precipitation is weak (as in stable precipitation), but decreasing (as in convective precipitation). As for accumulated snow, these graphs make it difficult to identify any dependence on the two parameters; however, see the next paragraph.

The variance-based SA offers a more objective assessment of these relationships. The left column in Fig. 4 shows the boxplot (across 20 days) of the sensitivity measures $S_i$ and $S_{ij}$. Based on the $S_i$ boxplots (top left), prcpfrac has a larger effect on convective precipitation

than mixlen does. The same pattern exists for stable precipitation, although to a lesser degree (top right). The effect of the two parameters on total precipitation is comparable; the comparable size of the boxplots (bottom left) further suggests that the variability is sufficiently large so as to preclude any reliable conclusions regarding the true effect of these parameters on total precipitation across multiple days. For accumulated snow (bottom right), the pattern is reversed: mixlen has a stronger effect on snow accumulation than prcpfrac; but the relatively larger size of the boxplots (cf. those for convective precipitation), implies that there is larger variability in the sensitivities.

The interaction between the parameters is assessed through the $S_{ij}$ boxplots (Fig. 4). For all four forecast quantities, the distribution of the $S_{ij}$ is contained to small values; the medians of the distributions are in the 0.02–0.09 range. In other words, about 2%–9% of the variability in $y$ can be attributed to interaction between prcpfrac and mixlen. A more detailed analysis of interactions is carried out in the next section.

## 5. Results: 11 parameters

Although the above scatterplots (e.g., Fig. 3) are useful for determining the sensitivity of forecast quantities on model parameters, that scatterplot-based approach is not possible when more than two parameters are at hand. In that situation, sensitivity can be assessed directly through the measures $S_i$ and $S_{ij}$. The distribution (across days and parameter values) of those measures is shown in Fig. 5, when the forecast quantity is convective precipitation (conv). With a median of $S_i$ around 40%, prcpfrac is the most influential parameter, followed by the $S_i \sim 25\%$ for temperature perturbation at the lifting condensation level in the convective parameterization (delt2KF). Recall that these percentages refer to the proportion of the variance of $y$ that can be attributed to the respective parameter. The remaining parameters have even lower contributions to the variability in convective precipitation.

The interaction between the parameters can be seen in the bottom panel of Fig. 5. The labels along the $x$ axis refer to the numerical labels assigned to each parameter (top-right corner of the figure). For example, the label "2, 3" refers to the parameters 2 and 3 [i.e., the cloud-radius factor in Kain–Fritsch (cloudrad) and prcpfrac]. Similarly, the label "3, 11" denotes the parameters 3 and 11 [i.e., prcpfrac and the slope intercept parameter for snow in the microphysics (snowsi)]. The horizontal dashed line at about $S_{ij} = 0.025$ marks the 95th percentile of the distribution of *all* $S_{ij}$ values—itself displayed at the leftmost side of the figure. As discussed in
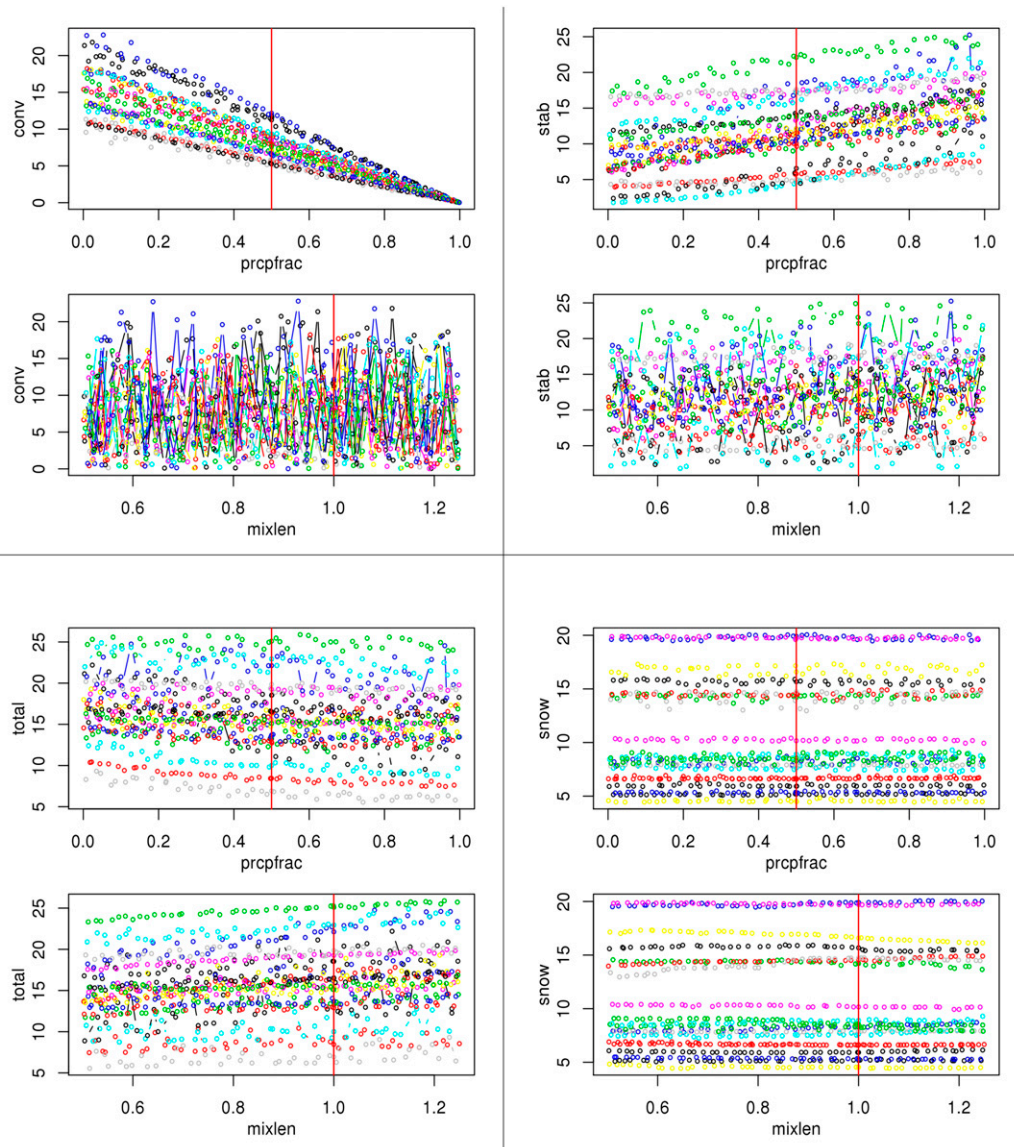
FIG. 3. Scatterplots showing the relationship between the forecast quantities (top left) convective precipitation, (top right) stable precipitation, (bottom left) total precipitation, and (bottom right) accumulated snow vs prcpfrac and mixlen. The different colors represent 20 different days. The vertical bars mark default values of the model parameters.

the method section, assuming the number of significant interactions is relatively small compared to the total number of interactions (here 55), then this distribution serves as an approximation to the null distribution of $S_{ij}$. In other words, if the distribution of a $S_{ij}$ is well above the 0.025 line, then one may conclude that the corresponding interaction is statistically significant. Conversely, if the 0.025 line falls well within the boxplots of $S_{ij}$, then one cannot conclude anything. And if the boxplot is well under the 0.025 line, then, the corresponding interaction is consistent with a no-interaction hypothesis.

An alternative method for assessing statistical significance is considered in the summary and discussion section.

For the pattern of distributions shown in the bottom panel of Fig. 5, the aforementioned assumption (that the number of large interactions is small) is not violated, because in fact no interactions appear to be anomalously large. In fact, the majority of the 55 boxplots fall well under the horizontal line at 0.025, suggesting that there is no statistically significant interaction between the corresponding parameters.
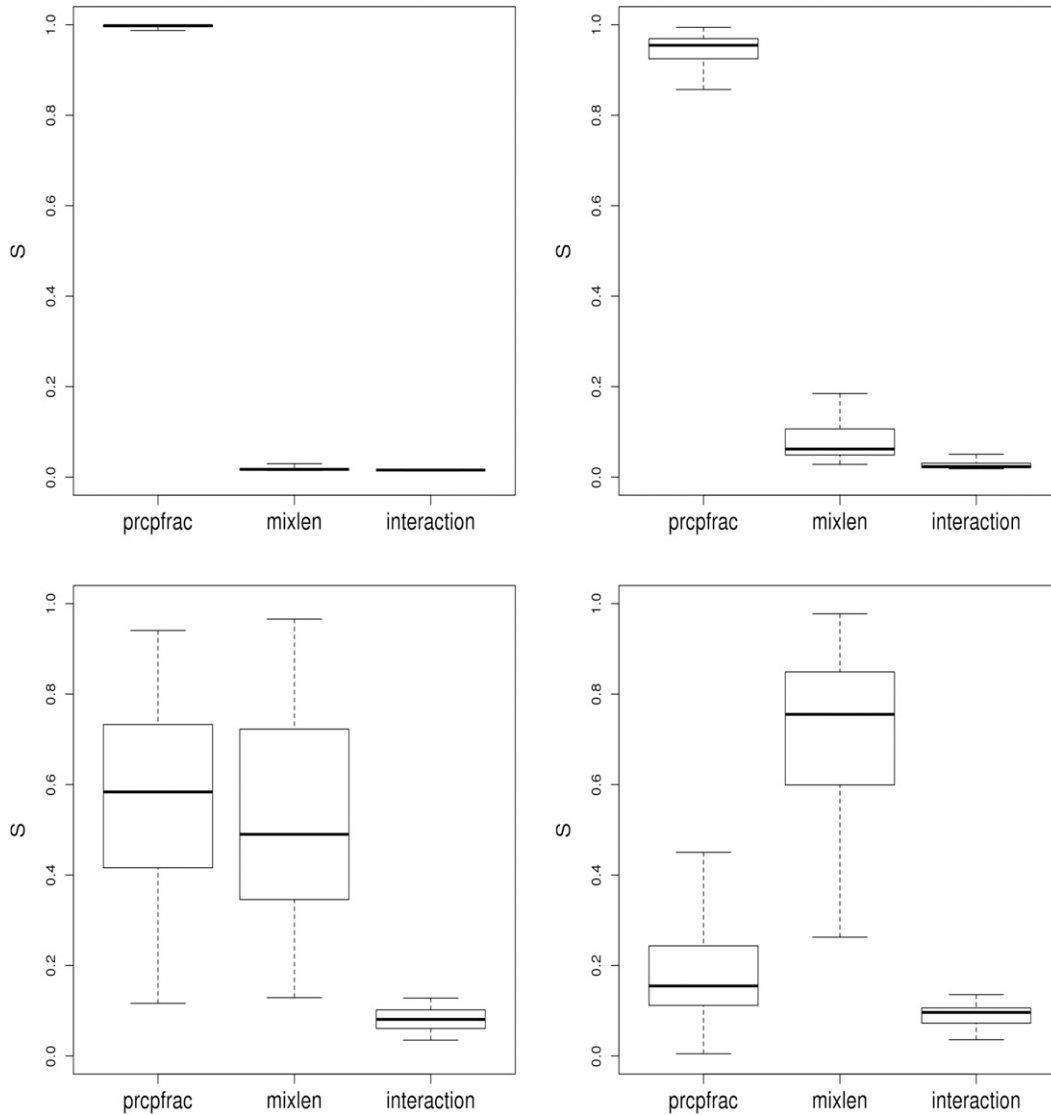
FIG. 4. The distribution of (top figure in each panel) sensitivity measures $S_i$ and (bottom figure in each panel) the interaction measure $S_{ij}$ for (top left) convective precipitation, (top right) stable precipitation, (bottom left) total precipitation, and (bottom right) accumulate snow.

A few of the boxplots do extend vertically to the point of including/overlapping the 0.025 line. The most notable examples are all the interactions involving the parameter "3" (i.e., prcpfrac). It follows, therefore, that no conclusion can be drawn regarding the value of the corresponding interactions. These interactions can be said to be statistically nonsignificant. Stated differently, the large variability of these interactions makes it difficult to infer whether or not the interactions are zero. One specific interaction ("1, 7") is marginally above the horizontal line, and so, is marginally nontrivial.

Then, given that the overwhelming majority of the boxplots are mostly below the 0.025 line, one can say

that these interactions are not statistically significant. In other words, the results found here are consistent with a no-interaction hypothesis. Moreover, the magnitude of all interactions is relatively small (between 0% and 6%, and with only one interaction extending to 10%). For these reasons, when setting the values of these parameters, it may be reasonable to ignore their interactions. Also, the absence of interactions suggests that the inclusion of other parameters in the sensitivity analysis will not affect the results found here (unless the additional parameters are found to interact with those examined here.)

The sensitivity results for stable precipitation are shown in Fig. 6. The most influential parameter is the    F6
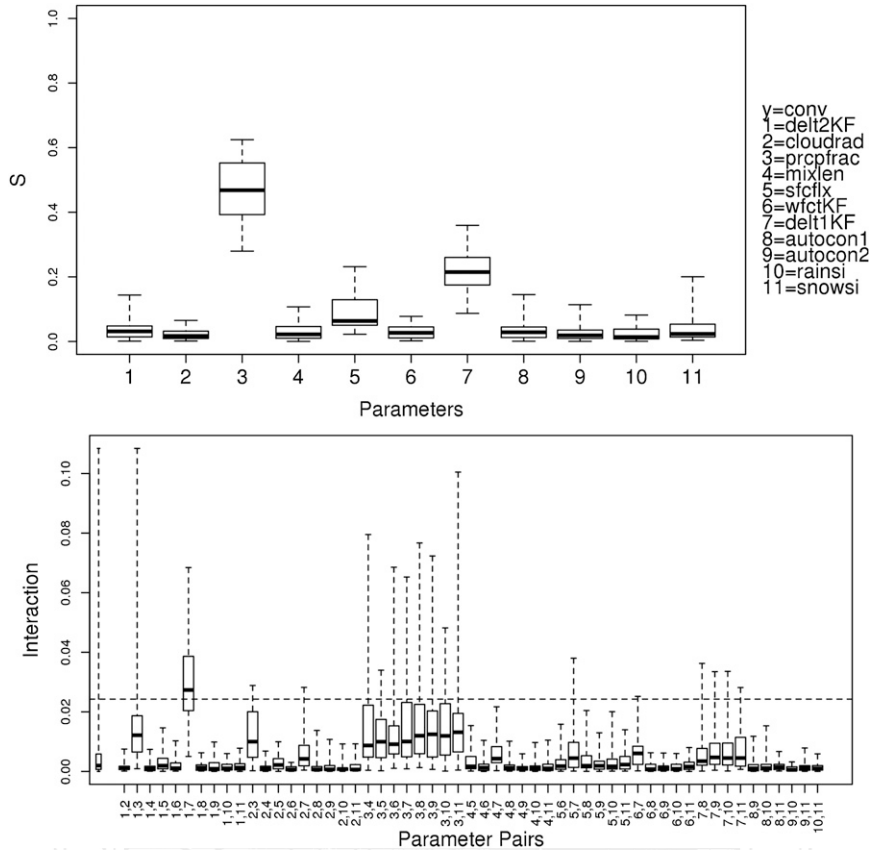
FIG. 5. (top) The distribution of the sensitivity measures $S_i$ and (bottom) the interaction measures $S_{ij}$ for convective precipitation (conv). See text for explanation of the labels along the $x$ axis.

linear factor that modifies the surface fluxes (sfcflx), explaining about 50% of the variability in stable precipitation. The variability of this quantity is, however, relatively large varying from 0% to 70%. The next most influential parameters are prcpfrac, delt1KF, and autocon2 (i.e., 3, 7, and 9). The remaining main effects are all small, never extending above and beyond explaining 10% of the variability in stable precipitation. As for interactions, all are either consistent with a no-interaction hypothesis (i.e., are mostly below the horizontal line), or have too much variability to allow for a definitive conclusion.

The results for total precipitation are shown in Fig. 7. Specifically, total precipitation is unambiguously effected most by sfcflx, and to a far lower degree by autocon2. The remaining parameters almost certainly have no effect on total precipitation. And the majority of the interactions, too, are almost certainly zero. The exceptions are the interactions involving parameter "5" (i.e., stcflx); the distributions of all of these interactions significantly overlap the horizontal line, and so their statistical significance cannot be established.

The sensitivity measures for accumulated snow are shown in Fig. 8. The most prominent feature of the main effects is that the two parameters that have relatively large main effects—sfcflx and snowsi—are also the ones with relatively large variability. This makes it difficult to "fine-tune" the accumulated snow. The majority of the interactions are again small or zero, with the exception of the interactions involving sfcflx and snowsi, whose value cannot inferred because of their large variability.

Figure 9 shows $E(y \mid x_i)$ for one day (1 January 2009), when $y$ represents convective precipitation. The 11 curves correspond to the 11 parameters under study. For visual acuity, the parameter values have been standardized to have zero mean and one standard deviation.[10] Most of the 11 curves are relatively "flat" and/or with small sensitivity measures. The thickest curve corresponds to prcpfrac, while the next thick curve corresponds to delt1KF. It can be seen that the expected value

---

[10] The gaps in the curves are a consequence of missing data (e.g., unavailability of the necessary files for running COAMPS).
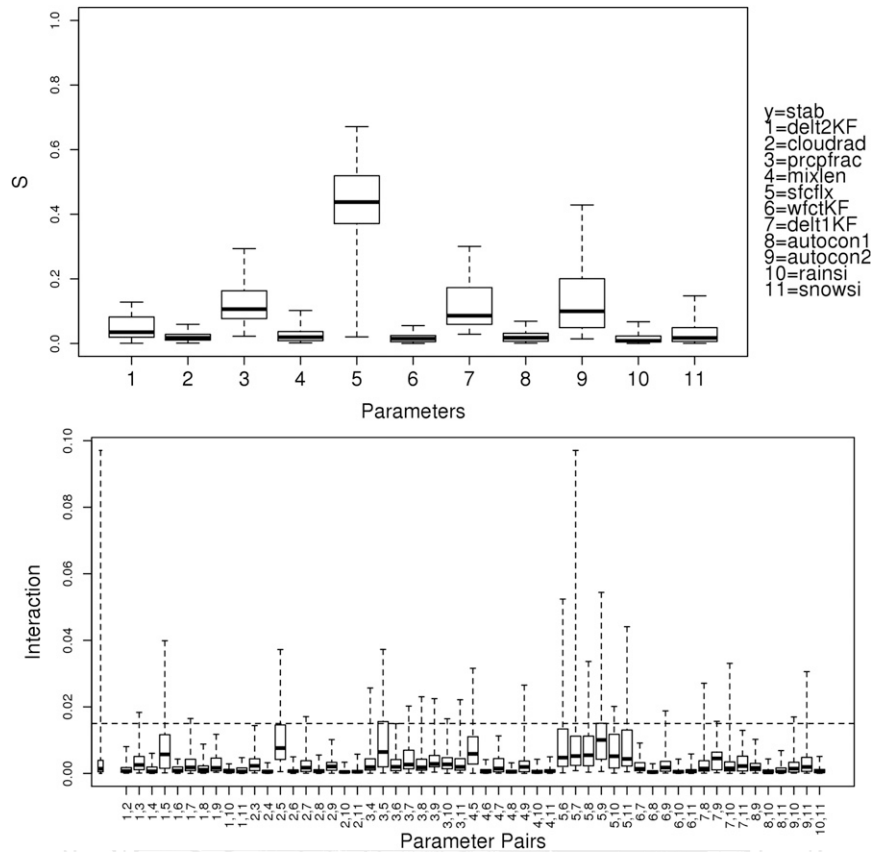
Monthly Weather Review (Proof Only)



FIG. 6. As in Fig. 5, but for stable precipitation (stab).

of convective precipitation decreases from about 15 mm to about 0, as prcpfrac spans its full range from 0% to 100%. The effect of delt1KF on convective precipitation is the opposite. These conclusions, although based on the $E(y\,|\,x_i)$ for one day, are consistent with the analysis of the sensitivity measures across all days in the study: most parameters have little or no effect on convective precipitation, with the exception of prcpfrac and delt1KF.

In this article, the focus has been on assessing the strength of main effects and interactions across all days and all parameters. For that reason, the variability displayed in the boxplots has reflected daily variability as well as sampling variability in parameter space. In some situations, however, it may be important to examine the two contributions separately–for example, in exploring the aforementioned possibility that interactions may exist only for certain days and/or only for certain pa- [F10] rameter values. Figure 10 shows the corresponding boxplots for when the forecast parameter is convective precipitation. It is interesting that the two sources of variability are comparable. This suggests that one should not be ignored in favor of the other, and that both play an important role in assessing sensitivity.

As mentioned previously, theoretical considerations suggest that LHS cannot give less precise estimates than SRS. This expectation was demonstrated in Marzban (2013) on the Lorenz'63 model. The application of the method to COAMPS leads to the same conclusion. Figure 11 compares the distribution of the sensitivity [F11] measure $S$, for the 11 parameters, under the two sampling schemes; the forecast quantity is convective precipitation; it can be seen that LHS provides more precise estimates (although only marginally) than SRS for parameters that have an effect on convective precipitation. For the parameters whose effect on convective precipitation is not clear, LHS and SRS yield comparable estimates.

## 6. Summary and discussion

In the approach adopted here, the assessment of the effect of model parameters on forecast quantities involves data generated from a computer experiment. Variance-based SA methods are ideally suited to analyzing such data. Recently, that method was illustrated on the Lorenz'63 model (Marzban 2011, 2013). Here, the method is applied to a complex, nonhydrostatic
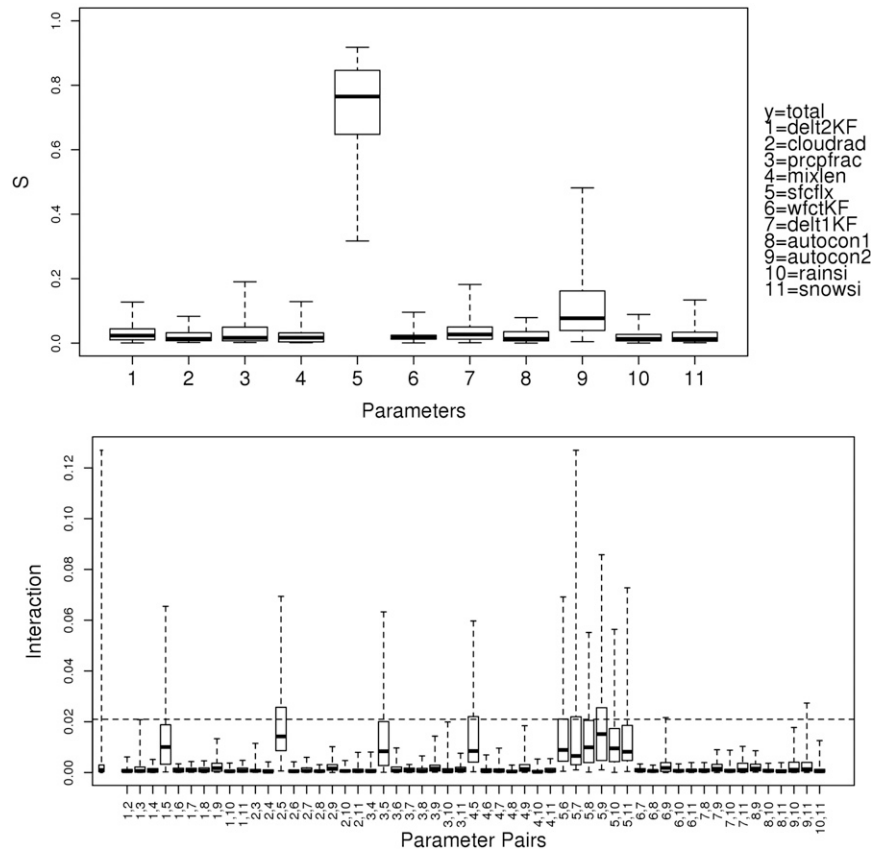
FIG. 7. As in Fig. 5, but for total precipitation (total).

mesoscale atmospheric prediction model with complex, nonlinear physical parameterizations (i.e., COAMPS). A total of 11 model parameters (Table 1) and 4 forecast quantities are considered: convective, stable, and total precipitation, and accumulated snow. The expectation that convective precipitation is most affected by prcpfrac is confirmed; delt1KF is also found to be a strong predictor. Stable and total precipitation have a similar set of main effects, with sfcflx appearing as the most influential parameter. The parameter sfcflx also has a strong influence on accumulated snow, but is second to snowsi. Although emphasis has been placed on identifying the influential parameters, it is worth pointing out that the identification of noninfluential parameters is also important, because one may then redirect resources to the fine-tuning of the influential parameters.

There appears to be little evidence for statistically or physically significant interactions between the parameters. As such, determining the "optimal" values of the parameters may be performed in simplistic (e.g., one at a time) fashion (Marzban 2013). However, a certain pattern of interactions is worth nothing: interactions involving a parameter with a large main effect are generally accompanied with large variability. For example, in Fig. 7, every interaction that involves sfcflx ("5")—a parameter with a large main effect—has a relatively large boxplot; some of these interactions are over 10% (e.g., "5, 6"). The same pattern exists for convective and stable precipitation (Figs. 5 and 6) and for accumulated snow (Fig. 8). It is possible that this association (between parameters with large main effects and interactions involving these parameters) is an artifact of the methodology. But another possible explanation is that there exist certain days and/or parameter values in the data for which the interactions are in fact significant. Work is under way to settle that question.

It is important to point out that these conclusions are true only for the specific settings of this study. For example, the conclusions may not hold if the resolution of the model is higher, if the measure of convective precipitation is different from the 90th percentile of precipitation forecasts, or if the model is run on dates beyond the span of those shown in Table 2.

Another brief comparison between the work done here and that of Holt et al. (2011) may provide further perspective. Holt et al. sample 29 points in the
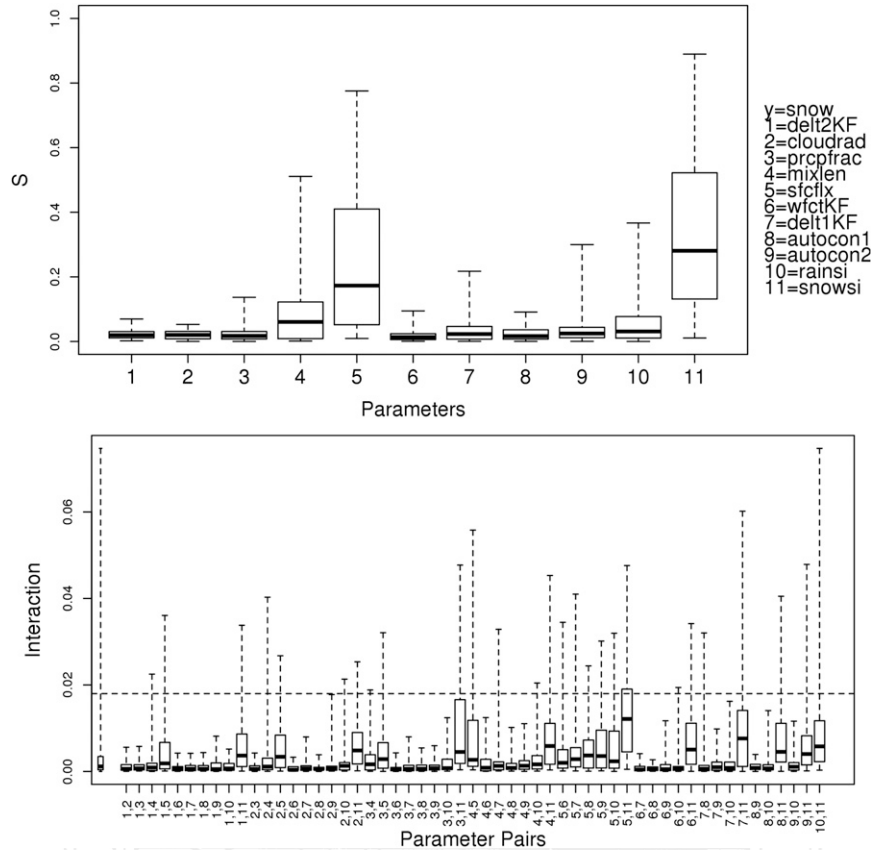
FIG. 8. As in Fig. 5, but for accumulated snow (snow).

11-dimensional parameter space, as opposed to the 99 points sampled here. The manner in which the points are selected is according to LHS and SRS, compared with the one-at-a-time (OAT) method used in Holt et al. The short-comings of the OAT method are explained in Marzban (2013). Beyond differences in sampling, another difference is that the variance-based method naturally provides scalar summary measures of sensitivity (e.g., $S$); Holt et al. produce all 29 "paths" and do not summarize them into a single sensitivity index. Finally, the current work examines daily variability, an issue that does not arise in Holt et al. at all.

It is possible to provide some physical insight into the results found here. Some of the results are readily interpreted from the description of the Kain–Fritsch (KF) parameterization scheme (Stensrud 2007). Referring to Fig. 3, it can be seen that convective precipitation is strongly and negatively correlated with prcpfrac [i.e., the fraction of available precipitation fed back to the grid from the KF scheme (also evident in Fig. 9), while stable or grid-scale precipitation is positively correlated with prcpfrac]. In other words, prcpfrac acts to regulate the proportion of precipitation going into convection (parameterized) versus going into stable

(explicit microphysics) precipitation, and, as shown in Fig. 3, with little or no impact on total precipitation. This is the strongest controlling factor on convective precipitation amount. The next strongest controlling factor between convective and stable precipitation is the convective trigger function parameter, delt1kf, which controls the temperature difference required to initiate convective precipitation. Again, there is a trade-off between convection and stable precipitation with little impact on total precipitation; however, with reverse correlations from prcpfac (i.e., higher delt1kf leads to stronger convection as expected). In contrast, sfcflx, the parameter controlling surface flux is positively correlated with all forms of precipitation—convective, stable, snow, and total—as it increases the amount of heat and moisture available. The other two significant factors, mixlen or PBL mixing length, and autocon2, an autoconversion factor for the microphysics are more obscure. The PBL mixing length affects the transfer of heat and moisture and the microphysics conversion factor impacts the autoconversion of cloud water.

The possible existence of interactions complicates both the understanding of the relationships, and what one does with that knowledge. In such situations, it is
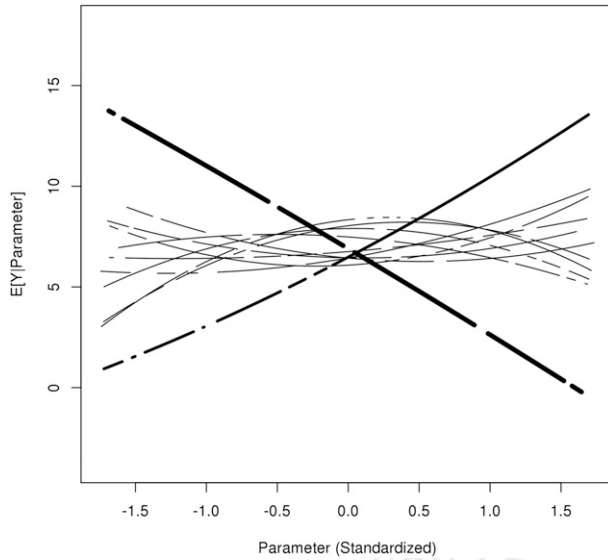
FIG. 9. The expected value of convective precipitation (conv) as a function of the 11 parameters (standardized to have zero mean and one standard deviation). The thickest curve corresponds to prcpfrac, and the next thickest curve corresponds to delt1KF—the parameters found to be most influential on conv, based on the variance-based method.
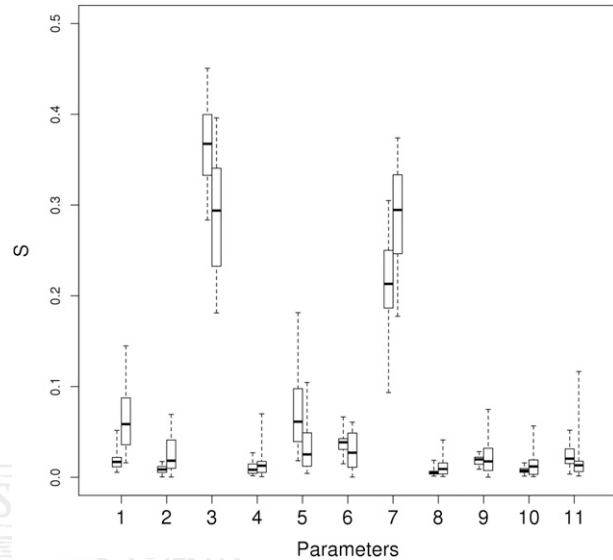
important to temper the results found here by another, independent method. One such method is based on regression. Specifically, for each day, and each parameter pair $(x_1, x_2)$, a regression model of the type $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$, is developed, where $y$ denotes the forecast quantity of interest. The term $\beta_{12}$ is a measure of the interaction between the two parameters, and its statistical significance can be tested via a $t$ test (Draper and Smith 1998). The distribution of the resulting $p$ values can then be used to assess statistical significance across multiple days. Under the null hypothesis of no interaction, the distribution of the $p$ values must be a uniform distribution between 0 and 1. Any deviation from uniformity can then be interpreted as evidence for rejecting the null hypothesis in favor of the alternative that an interaction exists. Figure 12 shows the distribution of the $p$ values when the forecast quantity is convective precipitation. The interactions for whom the $p$ values are mostly below the $\alpha = 0.05$ line (dashed horizontal line), cannot be uniform between 0 and 1, and are therefore potential candidates for significant interactions. According to the $p$ values in Fig. 12, there is one interaction that is statistically significant: "3, 7.". Although the results are not shown here, when this $p$-value-based approach is applied to the other forecast quantities, no such significant interactions are found. In short, the question of interactions is deserving of further work.

F12



FIG. 10. A comparison of (left) daily variability and (right) sampling variability in parameter space. The forecast quantity is convective precipitation.

Several other improvements to this work can be made. Here, regression models are used to estimate the conditional expected values. More sophisticated models, called Gaussian process regression models, have been proposed for that purpose (Rougier 2008). It is possible that such methods can better model the conditional expected values, with fewer points in parameter space. These models also allow for alternative ways of assessing statistical significance, and so, are currently under investigation.
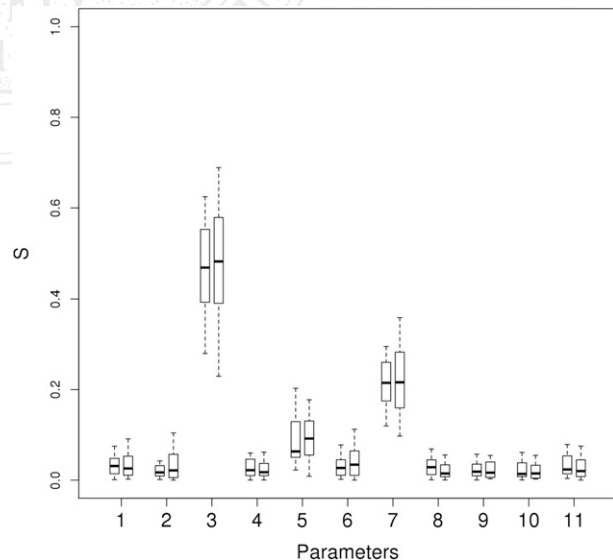


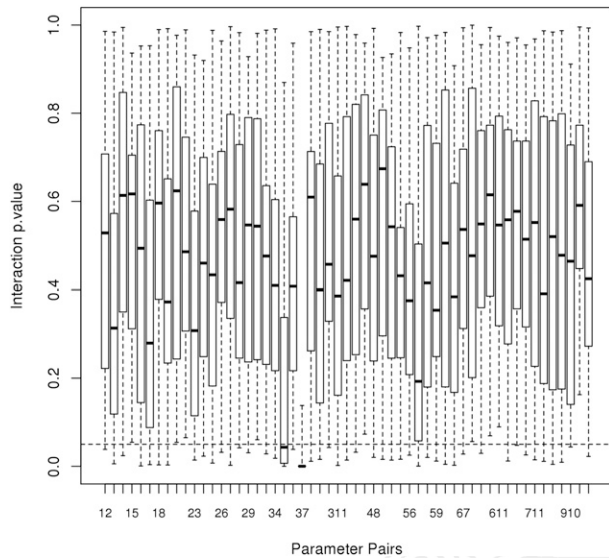FIG. 11. A comparison of (left) LHS and (right) SRS. The forecast quantity is convective precipitation.

FIG. 12. The distribution of regression-based *p* values testing the significance of each of the interactions. The horizontal line is at significance level $\alpha = 0.05$.

To gain a more complete understanding of what effects the forecast quantities, it is important to include more parameters in the analysis. With 11 parameters and 99 points in the experimental region, there has been no indication of overfitting in the development of the regression models. With more regression parameters, the dimension of the experimental region should be increased to avoid overfitting, but, at this time, it is unclear how much. It will be useful to perform simulation studies to obtain some sense of the relationship between the necessary number of points in parameter space and the number of parameters in the regression models.

Although the 36 days considered here span a significant portion of a year, the range is not sufficiently large to allow for a study of seasonal effects on the sensitivity of the parameters. Work is currently under way to expand the range to one full year, and then to multiple years.

The variance-based method used here involves only one output (forecast parameter). Revisions of the methodology allow for multiple outputs, simultaneously (Fasso 2006; Oakley and O'Hagan 2004). That approach will allow one to better model the effects and the interactions, because in practice one often sets the values of the model parameters while taking into account their effect on multiple forecast quantities.

Finally, for this study COAMPS was run with a grid spacing of 81 km. In future work, which will include the above-mentioned extensions, the runs will be performed at higher resolution. The computational demands will be significantly larger, but not formidable. Currently, a run

of COAMPS for a single set of 11 parameter values requires approximately 1 min on an 8-CPU computer. For 99 points in the empirical region, and over a span of 36 days, the run requires about 2.5 days ($99 \times 36$ min) to complete. Decreasing the grid spacing from 81 to 15 km (i.e., a factor of 5 increase in resolution) will require $5^3 = 125$ times longer. That length of time is unreasonably long on an 8-node computer, but is only 2.5 days ($2.5 \times 8/1000$) on a 1000-node supercomputer. Increasing the number of days in the analysis can also be compensated by taking them farther apart, which can actually be beneficial because the data across the different days will then be more independent. Given the feasibility of such experiments, it may even be possible to include "resolution" as one of the parameters in the sensitivity analysis.

## REFERENCES

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

Bolado-Lavin, R., and A. C. Badea, 2008: Review of sensitivity analysis methods and experience for geological disposal of radioactive waste and spent nuclear fuel. JRC Scientific and Tech. Rep., ?? pp., doi:10.2790/13786. [AU4]

Bowman, K. P., J. Sacks, and Y.-F. Chang, 1993: Design and analysis of numerical experiments. *J. Atmos. Sci.,* **50,** 1267–1278, doi: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, doi:10.1198/004017006000000453.

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, doi:10.1002/qj.1943.

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

Douglas, C. M., 2005: *Design and Analysis of Experiments.* John Wiley & Sons, 643 pp.

Doyle, J. D., Q. Jiang, R. B. Smith, and V. Grubii, 2011: Three-dimensional characteristics of stratospheric mountain waves during T-REX. *Mon. Wea. Rev.,* **139,** 3–23, doi:10.1175/2010MWR3466.1.

Draper, N. R., and H. Smith, 1998: *Applied Regression Analysis.* 3rd ed. Wiley-Interscience, 736 pp.

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

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. [Available online at http://www.iemss.org/iemss2006/sessions/all.html.]

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

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, doi:10.1111/j.1600-0870.2010.00505.x.

Harshvardhan, R. Davies, D. Randall, and T. Corsetti, 1987: A fast radiation parameterization for atmospheric circulation models. *J. Geophys. Res.,* **92,** 1009–1015, doi:10.1029/JD092iD01p01009.

Hodur, R. M., 1997: The Naval Research Laboratorys Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS). *Mon. Wea. Rev.,* **125,** 1414–1430, doi: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, doi:10.1007/s10236-011-0449-9.

Jiang, Q., and J. D. Doyle, 2009: The impact of moisture on mountain waves. *Mon. Wea. Rev.,* **137,** 3888–3906, doi:10.1175/2009MWR2985.1.

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.

Klemp, J., and R. Wilhelmson, 1978: The simulation of three-dimensional convective storm dynamics. *J. Atmos. Sci.,* **35,** 1070–1096, doi:10.1175/1520-0469(1978)035<1070:TSOTDC>2.0.CO;2.

Louis, J. F., 1979: A parametric model of vertical eddy fluxes in the atmosphere. *Bound.-Layer Meteor.,* **17,** 187–202, doi:10.1007/BF00117978.

Marzban, C., 2011: Sensitivity analysis in linear and nonlinear models: A review. Preprints, *Ninth Conf. on Artificial Intelligence and Its Application to the Environmental Sciences,* Seattle, WA, Amer. Meteor. Soc., 2A. [Available online at https://ams.confex.com/ams/91Annual/webprogram/Paper187333.html.]

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

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** (2), 239–245.

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

Rougier, J., 2008: Efficient emulators for multivariate deterministic functions. *J. Comput. Graph. Stat.,* **17,** 827–843, doi:10.1198/106186008X384032.

Rutledge, S. A., and P. V. Hobbs, 1983: The mesoscale and microscale structure of 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, doi:10.1175/1520-0469(1983)040<1185:TMAMSA>2.0.CO;2.

Sacks, J., S. B. Schiller, and W. J. Welch, 1989a: Designs for computer experiments. *Technometrics,* **31,** 41–47, doi:10.1080/00401706.1989.10488474.

——, W. J. Welch, T. J. Mitchell, and H. P. Wynn, 1989b: Design and analysis of computer experiments. *Stat. Sci.,* **4,** 297–437, doi:10.1214/ss/1177012413.

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, doi: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, 299 pp.

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

Stein, U., and P. Alpert, 1993: Factor separation in numerical simulations. *J. Atmos. Sci.,* **50,** 2107–2115, doi:10.1175/1520-0469(1993)050<2107:FSINS>2.0.CO;2.

Stensrud, C. D., 2007: *Parameterization Schemes: Keys To Understanding Numerical Weather Prediction Models.* Cambridge University Press, 457 pp.

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

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, doi:10.2307/1269548.