Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-273
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 13 November 2017

Geoscientific Model Development
Discussions
Open Access

On the Effect of Model Parameters on Forecast Objects

Caren Marzban[1,2]*, Corinne Jones[2], Ning Li[2], Scott Sandgathe[1]

[1] Applied Physics Laboratory

[2] Department of Statistics

Univ. of Washington, Seattle, WA 98195 USA

ABSTRACT

Many physics-based numerical models produce a gridded, spatial field of forecasts, e.g., a temperature "map." However, the field for some quantities such as precipitation generally consists of spatially coherent and disconnected "objects." Certain features of these objects (e.g., number, size, and intensity) are generally of interest. Here, a methodology is developed for assessing the impact of model parameters on features of forecast objects. Although, in principle, the objects can be defined by any means, here they are identified via clustering algorithms. The methodology is demonstrated on precipitation forecasts from a mesoscale numerical weather prediction model.

*Corresponding Author: marzban@stat.washington.edu

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-273
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 13 November 2017

1. Introduction

Complex, physics-based numerical models of natural phenomena often have parameters - henceforth, model parameters - whose values are generally not *a priori* specified. In such situations it is important to infer the manner in which the model parameters affect the outputs of the model (i.e., forecasts, or predictions), and often the techniques of Sensitivity Analysis (SA) are employed to assess the effects. There is a wide range of techniques from relatively simple one-at-a-time method (also known as the Morris method) where each model parameter is varied individually (e.g., Yu et al. (2013)), to multivariate approaches motivated by statistical methods of experimental design (Montgomery 2009) where the values of the model parameters are varied according to some optimization criterion. Alternative approaches can be found in Backman et al. (2017) where algorithmic differentiation is used, and in Kalra et al. (2017) where the underlying physics equations are integrated using quadrature methods. And yet another alternative is the adjoint method, commonly used in meteorological circles (Errico 1997).

It is difficult to classify these methods into a simple taxonomy (Bolado-Lavin and Badea 2008), but the terms Local and Global have been used to denote two broad categories (Saltelli et al. 2010, 2008); generally, local methods employ some sort of derivative of the model output with respect to inputs, while global techniques rely on a decomposition of the variance of the output in terms of the variance explained by the inputs. Comparisons of the various approaches are not common-place, because each approach is usually suited for a specific application where other methods may not be practically feasible. However, an example of the comparison of one global approach and one local (adjoint) approach on the

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-273
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 13 November 2017

Lorenz '63 model (Lorenz 1963) has been performed by Marzban (2013).

Another possible classification criterion is based on the purpose of the SA. Some SA work is performed for assessing how model parameters impact the model itself, not as a means to some other goal. For example, Lucas et al. (2013) uses a global SA method to explore the effect of model parameters on the probability of model crashes. By contrast, sometimes SA is performed as an intermediate step to another goal, such as the calibration of the model (Safta et al. 2015; Hacker et al. 2011; Laine et al. 2012; Ollinaho et al. 2014). All of these classification criteria are imperfect, as there exist works which fall "between" Global versus Local, or SA-only versus SA-for-calibration; some examples include Roebber (1989); Roebber and Bosart (1989); Robock et al. (2003). The work reported here falls into the Local and SA-only category; as such, although the proposed methodology can be used for calibration, no attempt is made to do so here.

In many SA studies, the output of the model (i.e., the response variable in the SA) is usually a single or a handful of scalar quantities. But there are situations in which the output is a gridded spatial field, e.g., temperature forecasts over a spatial region. Every grid point reflects a forecast at that location, and for a quantity like temperature the field as a whole has a smooth, continuous nature. SA is more complicated for precipitation fields, where the model output is a quantity whose spatial structure is not smooth and/or continuous. Indeed, there may be a coherent set of grid points that receive no precipitation at all, while an adjacent set of grid points will reflect a complex pattern of precipitation. In short, the spatial field of such quantities will contain "objects" within which precipitation does occur, surrounded by regions of little or no precipitation.

For such discrete fields, the assessment of the quality of the forecasts has given rise

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-273
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 13 November 2017

to a wide range of specialized techniques generally referred to as spatial verification (or evaluation) (Ahijevych et al. 2009; Baldwin et al. 2001, 2002; Brown et al. 2002; Casati et al. 2004; Davis et al. 2006a,b; Du and Mullen 2000; Ebert 2008; Ebert and McBride 2000; Gilleland et al. 2009; Hoffman et al. 1995; Keil and Craig 2007; Marzban and Sandgathe 2006, 2008; Marzban et al. 2008, 2009; Nachamkin 2004; Roberts and Lean 2008; Wealands et al. 2005; Wernli et al. 2008; Venugopal et al. 2005). A subset of these methods employs the notion of an object explicitly. In some applications, the object is defined subjectively - for example, by human experts. In other applications statistical methods for clustering (Everitt 1980) are used to identify/define objects within the field (Marzban and Sandgathe 2006, 2008). This clustering approach, which has been re-examined by Lakshmanan and Kain (2010), and more recently by Wang et al. (2015), is the basis of the object-identification procedure used in the present work.

Although no spatial verification/evaluation is done here, the importance of objects within the forecast field, and the development of clustering techniques for identifying them, calls for a SA framework wherein one can assess the effect of model parameters on the objects. In meteorology certain features of the clusters/objects are of special interest; they include size, location, intensity, and shape. Also, the assessment of sensitivity is highly intertwined with that of statistical significance. As such, the methodology developed here can be viewed as a SA with a multivariate response, wherein one can assess the impact (both the magnitude and the statistical significance) of model parameters on object features.

The model employed to demonstrate the methodology is COAMPS® (Hodur 1997), for which some SA work has already been done. Doyle et al. (2011) and Jiang and Doyle (2009) examine the effect of model parameters on mountain waves. Motivated by the work of Holt

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-273
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 13 November 2017

Geoscientific Model Development Discussions
Open Access
EGU

et al. (2011) who studied the effect of 11 model parameters on various characteristics of the forecasts, Marzban et al. (2014) used a global (variance-based) SA to study the effect of the same parameters and their interactions on mean (across the forecast domain) precipitation, and the center-of-gravity of precipitation.

2. Method

The methodology described in this paper involves two other techniques developed previously by some of the authors of this paper. In one, cluster analysis is used for identifying objects (Marzban and Sandgathe 2006, 2008; Marzban et al. 2008, 2009); in the other, SA is performed to assess the effect of model parameters on non-spatial features (e.g., domain mean) of the forecast field (Marzban et al. 2014). This section describes these components, puts forth the SA model, proposes means of assessing sensitivity and statistical significance, and describes the data used to demonstrate the methodology.

a. Data

The inputs of the numerical model examined here are 11 model parameters, and the outputs are forecasts of precipitation at each of 45×72 grid points, with a spacing of $81km$, covering the entire continental US, including coastal regions, and portions of Canada and Mexico. The SA method developed here requires data - technically, *computer data* - which are created by generating an ensemble (or sample) of inputs values, assimilating surface observations, and then running the model forward to produce 24h forecasts of precipitation

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-273
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 13 November 2017

amount at each grid point. As such, the SA results are contingent on the nature of this data, and consequently, care must be taken in the data-generation step of the methodology.

In order to include a wide range of weather phenomena, the data include 120 days from February 16 through July 2, 2009. Confirmed by visual examination of all 120 forecasts, this temporal period includes a comprehensive series of midaltitude synoptic systems traveling across the northern portion of the domain. These synoptic systems extend down into the southeastern US early in the period and are replaced by subtropical convective systems in the late spring and summer months. This subtropical activity also occurs in the southwestern portion of the domain (west coast of Mexico) during June and July in association with the southwest monsoon. The only apparent atypical weather appears to be a greater amount of convective activity off the east coast of the US associated with quasi-stationary or slow moving frontal systems during the period.

It is important that the data cases are as independent as possible. To that end, the 120 days are sampled at 3-day intervals in order to minimize temporal dependency, leading to 40 days for the analysis.

For each of the 40 days, 99 different values for 11 parameters are generated by Latin Hypercube Sampling (LHS). Said differently, for each day, a sample of size 99 is taken from the 11-dimensional space of the model parameters. This so-called "space-filling" sampling scheme assures 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 more precise estimates (at least, no less-precise estimates) than alternative sampling schemes (Cioppa and Lucas 2007; Montgomery 2009; Marzban 2013). LHS is appropriate when the model parameters are all continuous quantities (i.e., taking values on the Real line). For discrete or categorical inputs, Latin Square Designs

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-273
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 13 November 2017

(LSD) or Fractional Factorial Designs (FFD) can be employed to produce optimal samples (Montgomery 2009); these cases will be demonstrated in a separate article.

The 11 model parameters are shown in Table 1; the choice of these parameters is explained in Holt et al. (2011). As mentioned in that paper, these parameters were chosen for their anticipated sensitivity (through model tests and discussions with developers) of the parameterizations in an effort to chose parameters most likely to produce changes in the model output precipitation fields. Also, to focus on heavy precipitation, only the grid points whose convective precipitation amount exceeds the 90th percentile of precipitation across the domain are analyzed.

A very similar data set is used in Marzban et al. (2014) to assess the sensitivity of the average and center-of-gravity of precipitation (across the domain) with respect to the model parameters. Here, however, the precipitation fields are first subjected to cluster analysis (Sect. 3d), and then six cluster features (Sect. 3e) are employed as response variables in a multivariate SA.

b. Statistical Model

The SA methodology in (Marzban et al. 2014) is a variance-based approach which allows one to identify linear or nonlinear relationships between the forecast quantities and the model parameters, and even interactions between the model parameters. As a first approximation, however, it is sufficient to estimate only the linear (i.e., main) effects, because nonlinear and interaction effects are often much smaller than main effects; see pages 192, 230, 272, 314, 329 in Montgomery (2009), and pages 33-34 in Li et al. (2006). For this reason a linear

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-273
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 13 November 2017

regression-based model is adequate. Specifically, the effects of the model parameters are assessed via the least-squares estimate of the regression coefficients β_i in

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{11} x_{11} + \delta \ , \tag{1}$$

where x_i denote standardized model parameters, y is the forecast quantity (e.g., some cluster feature), and δ represents any source of variability in y other than from the model parameters. This linear model is further justified by the results (shown below) because when it is specialized to the case of one cluster (i.e., the entire spatial domain), it reproduces the results of the variance-based approach reported in Marzban et al. (2014).

There exists a realization of Eq. (1) in which the response is vector-valued; the model is called Multivariate Multiple Regression (MMR), wherein Eq. (1) is understood as a vector equation, where y, α, and β_i are all vectors. Ideally one could allow each component of the response vector to represent a forecast feature of a given object. However, the number of objects/clusters varies across the 99 values of the parameters and across days in the data. Methods for estimating MMR coefficients when the number of responses is a random variable (varying across cases) are not readily available. Therefore, for each of the six features, we consider three summary measures: The minimum, median, and maximum (across the clusters in the domain) of the feature. These three quantities can be thought of as a 3-point summary of the distribution of the feature, and they serve as the three responses in MMR.

The median across clusters is useful, because one can assess the effect of the model parameters on a "typical" cluster; minimum and maximum across clusters are useful because they allow one to assess whether a model parameter has an effect on **any** of the clusters in a field. For example, if it is found that a particular model parameter is positively (negatively)

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-273
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 13 November 2017

associated with the minimum (maximum) size across clusters, then one can conclude that the size of at least one of the clusters in the field is affected by that parameter. This is an important consideration, because if the size of at least one of the clusters is not affected by a parameter, then that parameter can be said to have no effect on the size of clusters. Additionally, consideration of the three summary measures, together, allows one to assess the effect of the model parameters on the distribution of the features.

The data on the response variables y are log-transformed to assure more bell-shaped histograms; this transformation is not necessary, but is useful when the regression coefficients are subjected to statistical tests, because many such tests assume relatively bell-shaped distributions.

c. Significance Tests

Testing the coefficients in the MMR model involves performing a large number of statistical tests ($40 \times 11 \times 6 \times 3$): one on each of 40 days, for each of 11 parameters, for each of six cluster features, and for each of three summary measures across clusters. A large number of tests, in turn, leads to an exponential growth in the probability of making *some* Type I error - a fact known as the multiple hypothesis testing problem (Montgomery 2009). A standard procedure in statistics for taming Type I errors is to divide the task into two stages (Montgomery 2009). In the first stage, one performs a single, often-called omnibus, hypothesis test of whether any of the parameters have an effect on any of the responses. In the present application, such a test reduces the number of tests to 40×6. If the null hypothesis cannot be rejected, then one performs no more tests, and the conclusion of the

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-273
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 13 November 2017

Geoscientific Model Development
Discussions
Open Access

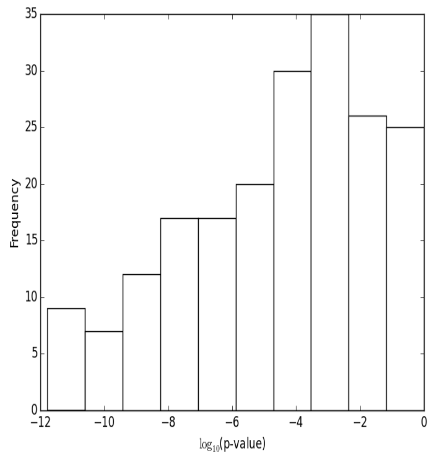


Fig. 1. Histogram of p-values from the multivariate tests across all days and response variables.

analysis is that there is no evidence that any of the parameters have an effect on any of the responses. If, however, the null hypothesis is rejected, then one proceeds to the second stage, i.e., testing each of the 40×6 effects, separately.

For the first stage, omnibus tests are readily available within MMR models (DelSole and Yang 2011; Fox et al. 2013; Rencher and Christensen 2012). Here, these tests were performed, yielding extremely small p-values, i.e., highly significant results (see Fig. 1), necessitating the second stage analysis.

For the second stage, a number of methods have been developed, again for the purpose of taming Type I errors; two of the more commonly employed methods are due to Tukey and Dunnet (Montgomery 2009). But these tests are generally complex procedures which in the end still involve a simplistic comparison of a p-value with a prespecified significance level. Although sufficient for hypothesis testing, these p-values provide no information on

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-273
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 13 November 2017

the magnitude of the effect. For this reason, instead, we adopt the more qualitative approach of examining the boxplot of the estimated regression coefficients directly.

The boxplots (shown in the next section) are generated and analyzed as follows. For each of the six cluster features, the response vector y is set to the minimum, median, and maximum (across clusters in the whole field) of that feature. For each of these three response variables, boxplots of the regression coefficients for the 11 model parameters are produced. The degree of overlap between each boxplot and the number zero reflects a visual (though qualitative) assessment of both the statistical significance and the magnitude of the effect of the corresponding model parameter on the response: If zero is well within the span of the boxplot, then one cannot conclude anything regarding the effect; if the boxplot is significantly above (below) zero, then one can conclude that the corresponding parameter has a positive (negative) effect on the response in question; and in such a case, the "distance" of the boxplot relative to zero provides a visual indication of the magnitude of the effect.

d. Cluster Analysis

There exists a wide range of clustering methods, each with their respective parameters (Everitt 1980). At one extreme, there exists a class of clustering methods wherein the desired number of cluster, NC, is specified by the user. A proven example in this class is called Gaussian Mixture Model (GMM) clustering (McLachlan and Peel 2000). At the other extreme, there exist clustering routines where NC does not play a role at all. One such method is called Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al. 1996). DBSCAN has two parameters, here denoted ϵ and min_samples. Roughly

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-273
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 13 November 2017

speaking, ϵ is the maximum distance between two grid points in order for them to be in the same cluster, and min_samples is the minimum number of grid points necessary to form a cluster.

These two approaches are selected here for demonstration because they allow for two very different ways in which a user can inject *a priori* knowledge into the analysis. For example, in some applications it may be more natural to specify the number of clusters, in which case GMM is a natural choice. On the other hand, DBSCAN is more natural if the user has knowledge of the typical size and distance between clusters. For example, consider a situation wherein the grid-spacing is relatively large (as is the case in this paper, i.e., $81km$), allowing one to examine only large scale precipitation. Although time of year and location are also important, but if one were to focus only on winter months in, say, the Pacific Northwest, then it is reasonable to set ϵ to 3 or 4. By contrast, if one is considering jet streaks, e.g., where some maximum wind speed value is reached, then ϵ can be closer to 1. As for min_samples, 4 or 5 are reasonable values for both precipitation and jet streak events, at the model resolution used here.

In addition to the way in which the respective parameters are handled, another reason why these two clustering methods are used here is that they occupy two other extremes in the family of clustering algorithms: GMM clustering belongs to a class of model-based algorithms (Banfield and Raftery 1993; Fraley and Raftery 2002) common in statistics circles because they are conducive to performing statistical tests, while DBSCAN assumes no underlying model, and for this reason is often employed in machine learning applications. For the SA component of the methodology developed here, it is not necessary for the objects to be defined by these or any other clustering algorithm; the objects may be defined by any other criterion or even by human experts.

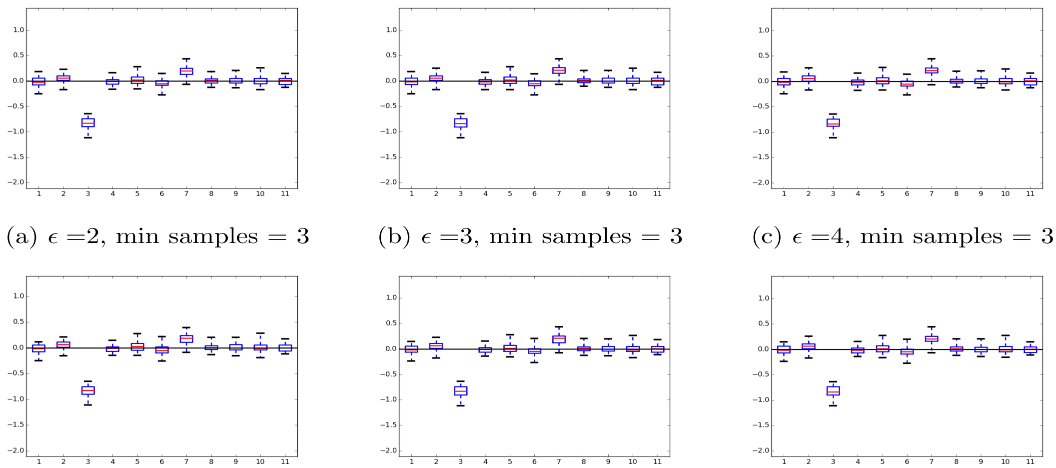
e. *Cluster Features*

In spatial verification some of the errors that are of interest include displacement, intensity, size/area, and shape error. The estimation of these errors presumes the ability to compute, respectively, the location, intensity, area, and shape of a cluster. Here, the latitude and longitude of the centroid of a cluster are taken as coordinates of its location; intensity is measured by the median (across the spatial extent of the cluster) of precipitation; and area is measured by the number of grid points in a cluster. The shape of a cluster, in GMM, is an ellipse, because that is the cross-section (i.e., level-set) of a bivariate Gaussian. Then, the eccentricity and orientation of the semi-major axis of the ellipse are natural for quantifying the shape of clusters. In DBSCAN, clusters are not restricted to have any specific shape. In order to be able to compare the two clustering algorithms, here an ellipse is "fitted" to the clusters, and again the eccentricity and orientation of the semi-major axis is used to represent the shape of the cluster. (Technically, the direction of the semi-major axis is defined to be the direction of the first eigenvector of the covariance matrix computed from the coordinates of all the grid points in a given cluster. The length of the semi-major axis is set to the largest eigenvalue.)

In short, the six cluster features examined here are latitude, longitude, intensity, area, orientation, and eccentricity. Also, recall that for each of these features, three summary measures are computed - minimum, median, and maximum - and used as the multivariate response vector in MMR (see Sect. 3b).

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-273
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 13 November 2017

Geoscientific Model Development
Open Access
Discussions



(a) ϵ =2, min samples = 3 (b) ϵ =3, min samples = 3 (c) ϵ =4, min samples = 3

(d) ϵ =2, min samples = 5 (e) ϵ =3, min samples = 5 (f) ϵ =4, min samples = 5

FIG. 2. Estimated regression coefficients (i.e. sensitivity of the model parameters) with median precipitation of the clusters as the response, after clustering with DBSCAN with various parameter values.

3. Results

As mentioned earlier, 40 forecasts are produced, each with 99 different values of 11 parameters in COAMPS. Each forecast field is clustered, and three summary measures (minimum, median, and maximum, all across clusters) are computed, each for six cluster features (latitude, longitude, etc.). First, an omnibus test is performed to test whether any of the 11 parameters have an effect on any of the three summary measures, on each day and for each cluster feature. Then, six MMR models are set up mapping the 11 parameters to three response variables. The daily variability - displayed as boxplots (e.g., Fig. 2 and Fig. 3) - for each of the regression coefficients offers a visual assessment of both the statistical significance and the magnitude of the effect of each parameter.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-273
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 13 November 2017

In the first stage of the analysis the response variable is a 3-dimensional vector, and an omnibus test is performed to test if any of the 11 parameters have an effect on any of the three response variables, for each day and each cluster feature. Such a test reduces the number of tests from $(40 \times 11 \times 6 \times 3)$ to $(40 \times 6) = 240$. The individual p-values are not shown here, but for DBSCAN their histogram is shown in Fig. 1. Evidently, all of the comparisons yield extremely small p-values. At a significance level of 0.05, out of the 240 tests, 29 p-values are not significant when using DBSCAN and 59 are not significant when using GMM. By examining the p-values, the majority of the non-significant results are associated with the tests when the response is the eccentricity of a cluster. If one applies the Bonferroni correction (Devore and Farnum 2005) to the significance level in order to account for the multiple tests, the significance level becomes $0.05/(40 \times 6) = 2 \times 10^{-4}$. At this significance level there are many more nonsignificant comparisons: 87 for DBSCAN and 94 for GMM. Upon making this correction, in addition to eccentricity some of the other features also emerge as being unaffected by any of the 11 parameters. Further details of these results are presented below.

Figure 2 shows the sensitivity results when the response is the median (across clusters) of precipitation intensity, and DBSCAN is employed with different parameters. The analogous results for GMM with different values of NC are not shown here, but they are similar. Recall that the variability displayed in each boxplot is due to the 40 days examined. First, note that all of the panels are mostly similar to one another, which implies that the sensitivity results are mostly unaffected by the parameters of the clustering algorithm.

It can also be seen that many of the 11 parameters have a histogram/boxplot of values mostly around zero. In other words, when considered across multiple days most of the 11 model parameters have no effect on the median of precipitation, The most obvious exception is parameter 3, which by virtue of having mostly negative values for its regression coefficient, is negatively associated with median precipitation. Parameter 7 not only has a weaker effect (because the median of the corresponding boxplot is closer to zero), it is also not as statistically significant (because zero falls well within the span of the boxplot). This parameter is positively associated with precipitation intensity in the typical (median) cluster, i.e., increasing the parameter leads to more intense clusters; more, below. All of these findings are consistent with those found for convective precipitation in Marzban et al. (2014) where a variance-based sensitivity was performed without any clustering at all. This consistency adds justification to the local/regression-based SA adopted here, i.e., Eq. (1).

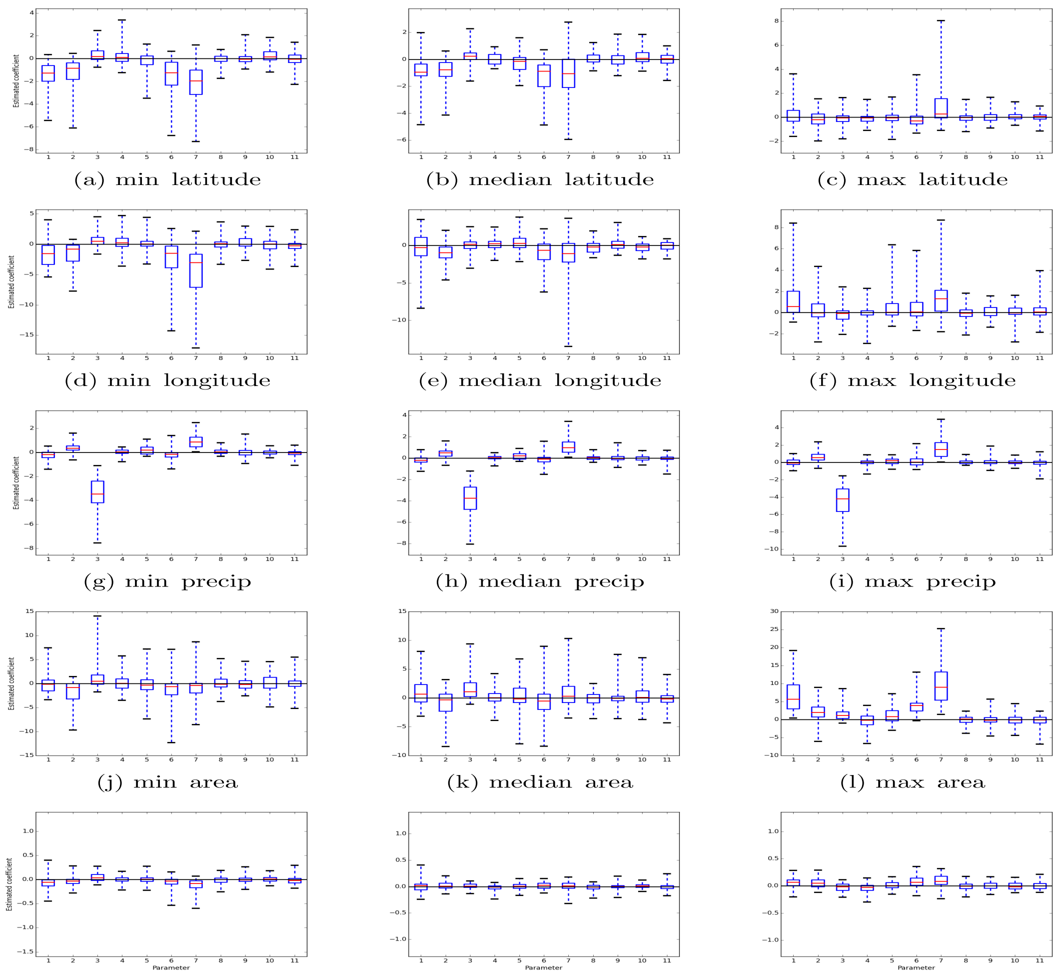
Figure 3 shows the effect of the model parameters on the latitude and longitude of the clusters (top two rows), amount of precipitation (middle row) in the clusters, and the area and orientation of the clusters (bottom two rows). The three columns correspond to the minimum, median, and maximum of a feature. Eccentricity has also been examined, but the results are not shown here because it is not affected by any of the 11 parameters; this conclusion is consistent with the results of the F-test performed in the first stage, mentioned above.

Examination of all of the panels suggests that parameters 4, 5, 8, 9, 10, 11 have little or no effect on any of the object features. By contrast, parameters 1, 2, 3, 6, and 7 appear to have varying effects depending on the object feature. Also, the orientation (in addition to eccentricity) of the clusters is unaffected by any of the parameters.

The strongest effects are from parameters 3 and 7 on the amount of precipitation. This relationship was already examined in Fig. 2; but now the same pattern can be seen in the

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-273
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 13 November 2017

Geoscientific Model Development Discussions



(a) min latitude (b) median latitude (c) max latitude
(d) min longitude (e) median longitude (f) max longitude
(g) min precip (h) median precip (i) max precip
(j) min area (k) median area (l) max area
(m) min orientation (n) median orientation (o) max orientation

FIG. 3. Estimated regression coefficients (i.e. sensitivity of the model parameters) on three summary measures (minimum, median, maximum) of different cluster features (latitude, longitude, amount of precipitation, and area and orientation of clusters. Eccentricity is not shown (see text). The clustering is done with DBSCAN with $\epsilon = 2\sqrt{2}$, min_samples = 3.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-273
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 13 November 2017

minimum, median, and maximum intensity (panels g, h, i in Fig. 3), which implies that the effect of parameters 3 and 7 is to shift down and up, respectively, the whole distribution of precipitation intensity.

The next strongest effects are those of parameters 1 and 7 on maximum area (panel l). Given that these two parameters have no effect on the minimum and median area (panels j and k), it follows that these parameters affect only the right tail of the distribution of size. In other words, by contrast to precipitation intensity whose distribution shifts when parameter 7 is varied, the distribution of size is stretched when that parameter changes. Parameter 6, too, appears to have an effect on maximum area, but to a lesser extent, both statistically and in magnitude.

Whereas parameter 1 tends to stretch out the distribution of area to the right, it appears to have the opposite effect on the minimum and median longitude of the clusters. The effect is weak in magnitude, but statistically significant. It does not affect the maximum longitude (panel f), and so, it stretches the distribution of longitude on the left, causing clusters to appear with smaller longitude, which given the encoding of the data used here, means to the west. Parameters 2, 6, and 7 appear to have the same effect as parameter 1.

The latitude appears to be weakly affected by some of the parameters. For example, parameter 7, and to a much lesser degree parameter 1, is positively associated with median and maximum latitude, but negatively associated with minimum latitude. In other words, increasing parameter 7 increases the width of the distribution of latitude values, causing them to be more spread out along the latitudes.

All of the above conclusions are based on clustering with DBSCAN with $\epsilon = 2\sqrt{2}$ and min_samples=3. To test the robustness of these results the same analysis was repeated but

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-273
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 13 November 2017

with GMM as the cluster algorithm and with $NC = 3$. The results (not shown here) are mostly the same. One relatively clear difference between the DBSCAN and GMM results is in the effect of parameters 1 and 7 on area; whereas with DBSCAN those parameters have an effect only on the maximum area, the results based on GMM suggest a significant effect on all three cluster features (minimum, median, and maximum area).

Further differences between DBSCAN and GMM sensitivity results are found when one performs a multivariate test for the effect of the model parameters across **all** days. For DBSCAN, the p-values corresponding to each of the six cluster features are all found to be nearly zero. So, some of the model parameters do have a significant effect on some of the features. The same is true for GMM, with the exception of latitude and eccentricity for which there is no evidence of an effect (p-values 0.435 and 0.290, respectively). It may appear that these results are contradictory, but they are not because the respective parameters of the two clustering algorithms have not been tuned to render them comparable. Specifically, the DBSCAN parameters are $\epsilon = 2\sqrt{2}$ and min_samples=3, while for GMM the parameter NC is set to three. In other words, the differences are due to the way in which the two clustering algorithms handle their respective parameters. As mentioned earlier, such differences do not point to defects in the methodology; they simply reflect the choice of what the user considers to be an object.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-273
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 13 November 2017

4. Conclusion and Discussion

It is shown that by employing methods of cluster analysis and sensitivity analysis one can assess the magnitude and statistical significance of the effect of model parameters on features (location, intensity, size, and shape) of objects within forecast fields. The framework also allows one to assess the impact of the model parameters on the distribution of forecast features. For example, one can reveal the model parameters that affect the overall location and/or width of the distribution of object features, and those which impact the shape of the distribution, e.g., by stretching out the left and/or right tail. The approach does not point to any "optimal" values of the model parameters, for that would require optimizing the model parameters to maximize some measure of agreement between forecasts and observations. In other words, although the work here lays the foundation for tuning the model parameters for the purpose of improving forecasts in terms of metrics that arise naturally in spatial verification/evaluation methods, no such tuning is performed here.

Given the novelty of the proposed framework, some recommendations are in order. The choice of the clustering algorithm depends on the specific user. Indeed, there are situations in which clusters/objects in a field are identified by human experts. For these reason, no specific clustering algorithm is recommended. A similar philosophy is adopted with respect to the values of the parameters of the clustering algorithms; they may be specified by the user, or varied across a range of values, depending on the specific application. Although there exist statistical criteria that lead to unique values for the parameters, the criteria involve the optimization of some other quantity, e.g., Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC). As such, the ambiguity in the choice of the clustering algorithm, or the values of their parameters, is simply replaced with the ambiguity of selecting the appropriate criterion. Therefore, again, no attempt is made to optimize the values of the parameters. It is assumed that the user has sufficient information about the underlying physics to either specify the number of physical objects (or a range thereof), or the typical size and distance between physical objects.

It is worth pointing out that at least in meteorology, it is not uncommon for different human experts to have different notions of an object in the forecast field. As such, the ambiguities discussed above are not specific to clustering algorithms, but are inherent to any object-based approach. In spite of this inherent ambiguity, many spatial verification techniques generally rely on some notion of an object. The main reason is that accounting for objects in a forecast field is a first step in the verification/evaluation process, and the manner in which objects are defined is of secondary importance.

While this paper is primarily about a methodology, it is worthwhile to provide a possible physical explanation for at least the strongest results in the COAMPS application. The strongest influence or sensitivity is from parameter 3, the fraction of available precipitation fed back to the grid from the Kain-Fritsch scheme. Increasing this fraction reduces convective precipitation and, based on the results in Marzban et al. (2014), increases stable precipitation, while not affecting total precipitation. It also is responsible for weakening the convective precipitation, i.e., increasing the number of weak systems. The next largest sensitivity is from parameter 7, which controls the temperature difference required to initiate convective precipitation. Again, as shown in Marzban et al. (2014), this parameter also controls a trade-off between convective and stable precipitation and has little effect on total precipitation (along with parameter 1). Parameters 1 and 7 do increase the area of convective precipitation in large precipitation events but not in smaller (areal) precipitation events, likely due to the trade-off between stable and convective precipitation in large events such as frontal systems and mesoscale clusters. This process may also explain the apparent increase in east-west areal coverage and the intensification of precipitation events, as found here.

Several generalizations of the proposed methodology are possible. In Marzban et al. (2008) it has been shown that clustering can be done not only in the 2-dimensional space of latitude and longitude of each grid point, but also in the 3-dimensional space that includes the amount of precipitation at each grid point. In fact, one may argue that the inclusion of more meteorological quantities in the clustering phase ought to lead to more meteorologically relevant objects being identified. In turn, this is more likely to lead to more realistic representation of the effect of the parameters on the object features. The object features may also be extended or revised. For example, here the shape of an object is approximated by an ellipse. But it is possible to use more sophisticated methods of shape analysis (Bookstein 1991; Lack et al. 2010; Micheas et al. 2007; Lakshmanan et al. 2009) to model more complex shapes. Another possible generalization is to allow for interactions between model parameters. Although the statistical model used here does account for covariance between the model parameters, and between the response variables, no explicit interaction is introduced. The inclusion of such terms is straightforward, and is unlikely to lead to overfitting, at least in linear models such as MMR.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-273
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 13 November 2017

5. Code and/or data availability

The code and the data analyzed here occupy about 4.0G of computer space, and are available upon request from the corresponding author, or from https://doi.org/10.5281/zenodo.1043542

6. Competing Interests

The authors declare that they have no conflict of interest

7. Acknowledgments

This work has received support from Office of Naval Research (N00014-12-G-0078 task 29) and National Science Foundation (AGS-1402895). The authors are grateful to James D. Doyle and Nicholas C. Lederer for providing invaluable support.

REFERENCES

Ahijevych, D., Gilleland, D. E., Brown, B. G., and Ebert, E. E.: Application of spatial verification methods to idealized and NWP-gridded precipitation forecasts, Wea. Forecasting, 24, 1485–1497, 2009.

Backman, J., Wood, C., Auvinen, M., Kangas, L., Hannuniemi, H., Karppien, A., and

Kukkonen, J.: Sensitivity analysis of the meteorological pre-processor MPP-FMI 3.0 using algorithmic differentiation, Geosci. Model Dev. Discuss., (in review), 2017.

Baldwin, M. E., Lakshmivarahan, S., and Kain, J. S.: Verification of mesoscale features in NWP models, in: Amer Meteor. Soc., 9th Conf. on Mesoscale Processes, pp. 255–258, Ft. Lauderdale, FL., 2001.

Baldwin, M. E., Lakshmivarahan, S., and Kain, J. S.: Development of an "events-oriented" approach to forecast verification, in: 15th Conf. Numerical Weather Prediction, San Antonio, TX, 2002.

Banfield, J. D. and Raftery, A. E.: Model-based Gaussian and non-Gaussian clustering, Biometrics, 49, 803–821, 1993.

Bolado-Lavin, R. and Badea, A. C.: Review of sensitivity analysis methods and experience for geological disposal of radioactive waste and spent nuclear fuel, JRC Scientific and Technical Report, Available online, 2008.

Bookstein, F. L.: Morphometric Tools for Landmark Data: Geometry and Biology, Cambridge, 1991.

Brown, B. G., Mahoney, J. L., Davis, C. A., Bullock, R., and Mueller, C.: Improved approaches for measuring the quality of convective weather forecasts, in: 16th Conference on Probability and Statistics in the Atmospheric Sciences, pp. 20–25, Orlando, FL, 2002.

Casati, B., Ross, G., and Stephenson, D.: A new intensity-scale approach for the verification of spatial precipitation forecasts, Met. App., 11, 141–154, 2004.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-273
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 13 November 2017

Cioppa, T. and Lucas, T.: Efficient nearly orthogonal and space-filling latin hypercubes, Technometrics, 49(1), 45–55, 2007.

Davis, C., Brown, B., and Bullock, R.: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas, Mon. Wea. Rev., 134, 1772–1784, 2006a.

Davis, C. A., Brown, B., and Bullock, R.: Object-based verification of precipitation forecasts. Part II: Application to convective rain systems, Mon. Wea. Rev., 134, 1785–1795, 2006b.

DelSole, T. and Yang, X.: Field Significance of Regression Patterns, J. Climate, 24, 5094–5107, 2011.

Devore, J. and Farnum, N.: Applied Statistics for Engineers and Scientists, Thomson, 2005.

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

Du, J. and Mullen, S. L.: Removal of Distortion Error from an Ensemble Forecast, Mon. Wea. Rev., 128, 3347–3351, 2000.

Ebert, E.: Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework, Meteor. Appl., 15 (1), 51–64, 2008.

Ebert, E. E. and McBride, J. L.: Verification of precipitation in weather systems: determination of systematic errors, Jour. Hydrology, 239, 179–202, 2000.

Errico, R. M.: What is an Adjoint Model?, Bull. Amer. Meteor. Soc, 78, 2577–2591, 1997.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of Knowledge Discovery and Data Mining (KDD-96), vol. 96(34), pp. 226–231, 1996.

Everitt, B. S.: Cluster Analysis, Heinemann Educational Books, London, 1980.

Fox, J., Friendly, M., and Weisberg, S.: Hypothesis tests for multivariate linear models using the car package, The R Journal, 5(1), 39–52, 2013.

Fraley, C. and Raftery, A.: Model-Based Clustering, Discriminant Analysis, and Density Estimation, Journal of the American Statistical Association, 97, 611–631, 2002.

Gilleland, D. E., Ahijevych, D., Brown, B. G., Casati, B., and Ebert, E. E.: Intercomparison of Spatial Forecast Verification Methods, Wea. Forecasting, 24, 1416–1430, 2009.

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

Hodur, R. M.: The Naval Research Laboratorys Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS), Mon. Wea. Rev., 125, 1414–1430, 1997.

Hoffman, R. N., Liu, Z., Louis, J.-F., and Grassotti, C.: Distortion representation of forecast errors, Mon. Wea. Rev., 123, 2758–2770, 1995.

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

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-273
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 13 November 2017

Jiang, Q. and Doyle, J. D.: The impact of moisture on Mountain Waves, Mon. Wea. Rev., 137, 3888–3906, 2009.

Kalra, T. S., Aretxabaleta, A., Seshadri, P., Ganju, N. K., and Beudin, A.: Sensitivity Analysis of a Coupled Hydrodynamic-Vegetation Model Using the Effectively Subsampled Quadratures Method, Geosci. Model Dev. Discuss., (in review), 2017.

Keil, C. and Craig, G. C.: A displacement-based error measure applied in a Regional Ensemble Forecasting System, Mon. Wea. Rev., 135(9), 3248–3259, 2007.

Lack, S. A., Limpert, G. L., and Fox, N. I.: An object-oriented multiscale verification scheme, Wea. Forecasting, 25(1), 79–92, 2010.

Laine, M., Solonen, A., Haario, H., and Järvinen, H.: Ensemble prediction and parameter estimation system: the method, Q. J. R. Meteorol. Soc, 138, 289–297, 2012.

Lakshmanan, V. and Kain, J. S.: A Gaussian Mixture Model Approach to Forecast Verification, Wea. Forecasting, 25(3), 908–920, 2010.

Lakshmanan, V., Hondl, K., and Rabin, R.: An Efficient, general-purpose technique for identifying storm cells in geospatial image, J. Atmos. Oceanic Technol., 26, 523–537, 2009.

Li, X., Sudarsanam, N., and Frey, D. D.: Regularities in data from factorial experiments, Complexity, 11(5), 32–45, 2006.

Lorenz, E. N.: Deterministic non-periodic flow, J. Atmos. Sci., 20, 130–141, 1963.

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

Marzban, C.: Variance-based Sensitivity Analysis: An illustration on the Lorenz '63 model, Mon. Wea. Rev., 141(11), 4069–4079, 2013.

Marzban, C. and Sandgathe, S.: Cluster analysis for verification of precipitation fields, Wea. Forecasting, 21(5), 824–838, 2006.

Marzban, C. and Sandgathe, S.: Cluster Analysis for Object-Oriented Verification of Fields: A Variation, Mon. Wea. Rev., 136, 1013–1025, 2008.

Marzban, C., Sandgathe, S., and Lyons, H.: An Object-oriented Verification of Three NWP Model Formulations via Cluster Analysis: An objective and a subjective analysis, Mon. Wea. Rev., 136 (9), 3392–3407, 2008.

Marzban, C., Sandgathe, S., Lyons, H., and Lederer, N.: Three Spatial Verification Techniques: Cluster Analysis, Variogram, and Optical Flow, Wea. Forecasting, 24(6), 1457–1471, 2009.

Marzban, C., Sandgathe, S., Doyle, J. D., and Lederer, N. C.: Variance-Based Sensitivity Analysis: Preliminary Results in COAMPS, Mon. Wea. Rev., 142, 2028–2042, 2014.

McLachlan, G. J. and Peel, D.: Finite Mixture Models, John Wiley & Sons, Hoboken, NJ USA, 2000.

Micheas, A. C., Fox, N. I., Lack, S. A., and Wikle, C. K.: Cell identification and verification of QPF ensembles using shape analysis techniques, J. Hydrology, 343, 105–116, 2007.

Montgomery, D. C.: Design and Analysis of Experiments, Wiley & Sons, 7th edition, 2009.

Nachamkin, J. E.: Mesoscale verification using meteorological composites, Mon. Wea. Rev., 132, 941–955, 2004.

Ollinaho, P., ärvinen, H., Bauer, P., Laine, M., Bechtold, P., Susiluoto, J., and Haario, H.: Optimization of NWP model closure parameters using total energy norm of forecast error as a target, Geosci. Model Dev., 7(5), 1889–1900, 2014.

Rencher, A. C. and Christensen, W. F.: Methods of Multivariate Analysis, John Wiley & Sons, Inc., Hoboken, NJ, USA, doi:10.1002/9781118391686.ch10, 2012.

Roberts, N. M. and Lean, H. W.: Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events, Mon. Wea. Rev., 136, 78–97, 2008.

Robock, A., Luo, L., Wood, E. F., Wen, F., Mitchell, K. E., Houser, P., Schaake, J. C., Lohmann, D., Cosgrove, B., Sheffield, J., Duan, Q., Higgins, R. W., Pinker, R. T., Tarpley, J. D., Basara, J. B., and Crawford, K. C.: Evaluation of the North American Land Data Assimilation System over the southern Great Plains during warm seasons, J. Geophys. Res., 108, 8846–8867, 2003.

Roebber, P.: The role of surface heat and Moisture Fluxes Associated with large-scale ocean current meanders in maritime cyclogenesis, Mon. Wea. Rev., 117, 1676–1694, 1989.

Roebber, P. and Bosart, L.: The sensitivity of precipitation to circulation details. part i: an analysis of regional analogs, Mon. Wea. Rev., 126, 437–455, 1989.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-273
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 13 November 2017

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

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

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., and Tarantola, S.: Variance based sensitivity analysis of model output: Design and estimator for the total sensitivity index, Computer Physics Communications, 181, 259–270, 2010.

Venugopal, V., Basu, S., and Foufoula-Georgiou, E.: A new metric for comparing precipitation patterns with an application to ensemble forecasts, J. Geophys. Res., 110, D8, D08 111 10.1029/2004JD005 395, 2005.

Wang, Y. H., Fan, C. R., Zhang, J., Niu, T., Zhang, S., and Jiang, J. R.: Forecast Verification and Visualization based on Gaussian Mixture Model Co-estimation, Computer Graphics Forum, 34, 99–110, 2015.

Wealands, S. R., Grayson, R. B., and Walker, J. P.: Quantitative comparison of spatial fields for hydrological model assessment: some promising approaches, Advances in Water Resources, 28, 15–32, 2005.

Wernli, H., Paulat, M., Hagen, M., and Frei, C.: SAL - A Novel Quality Measure for the Verification of Quantitative Precipitation Forecasts, Mon. Wea. Rev., 136, 4470–4487, 2008.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-273
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 13 November 2017

Yu, Y. Y., Finke, P. A., Wu, H. B., and Guo, Z. T.: Sensitivity analysis and calibration of a soil carbon model (SoilGen2) in two contrasting loess forest soils, Geosci. Model Dev., 6, 29–44, 2013.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-273
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 13 November 2017

ID	Name (Unit)	Description	Default	Range
1	delt2KF ($^{\circ}C$)	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 ($^{\circ}C$)	Another method to perturb the temperature at the LCL in KF	0	-2, 2
8	autocon1 ($\frac{kg}{m^3 s}$)	Autoconversion factors for the microphysics	0.001	1e-4, 1e-2
9	autocon2 ($\frac{kg}{m^3 s}$)	Autoconversion factors for the microphysics	4e-4	4e-5, 4e-3
10	rainsi ($\frac{1}{m}$)	Microphysics slope intercept parameter for rain	8.0e6	8.0e5, 8.0e7
11	snowsi ($\frac{1}{m}$)	Microphsyics slope intercept parameter for snow	2.0e7	2.0e6, 2.0e8

KF = Kain-Fritsch, PBL = Planetary Boundary Layer, LCL = Lifted Condensation Level

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