Meteorological Applications
The Royal Meteorological Society Journal for applied meteorologists, forecasters and users of meteorological services
RMetS Royal Meteorological Society

A Methodology for Sensitivity Analysis of Spatial Features in Forecasts: The Stochastic Kinetic Energy Backscatter Scheme

Journal:	*Meteorological Applications*
Manuscript ID	MET-16-0165.R2
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	25-May-2018
Complete List of Authors:	Marzban, Caren; University of Washington, Applied Physics Lab; Dept of Statistics, Tardif, Robert; University of Washington, Atmospheric Sciences Hryniw, Natalia; University of Washington, Atmospheric Sciences Sandgathe, Scott; University of Washington, Applied Physics Laboratory
Keywords:	Sensitivity Analysis < Modelling, Statistical Models < Modelling, NWP < Modelling
Manuscript keywords:	

A Methodology for Sensitivity Analysis of Spatial Features in Forecasts: The Stochastic Kinetic Energy Backscatter Scheme

Caren Marzban[1,2], Robert Tardif[3], Natalia Hryniw[3], Scott Sandgathe[1]

[1] Applied Physics Laboratory,
[2] Department of Statistics,
[3] Department of Atmospheric Sciences,
Univ. of Washington, Seattle, WA 98195 USA

Abstract

Stochastic Kinetic Energy Backscatter Schemes (SKEBS) are introduced in numerical weather forecast models to represent uncertainties related to unresolved subgrid-scale processes. These schemes are formulated using a set of parameters that must be determined using physical knowledge and/or to obtain a desired outcome. Here, a methodology is developed for assessing the effect of four factors on spatial features of forecasts simulated by the SKEBS-enabled Weather Research and Forecasting (WRF) model. The four factors include two physically motivated SKEBS parameters (determining amplitude of perturbations applied to streamfunction and potential temperature tendencies), a purely stochastic element (a seed used in generating random perturbations), and a factor reflecting daily variability. A simple threshold-based approach for identifying coherent objects within forecast fields is employed, and the effect of the four factors on object features (e.g., number, size, and intensity) is assessed. Four object types are examined: upper-air jet streaks, low-level jets, precipitation areas, and frontal boundaries. The proposed method consists of a set of standard techniques in experimental design, based on the analysis of variance, tailored to sensitivity analysis. More specifically, a Latin Square Design is employed to reduce the number of model simulations necessary for performing the sensitivity analysis. Fixed effects and random effects models are employed to assess the main effects and the percentage of the total variability explained by the four factors. It is found that the two SKEBS parameters do not have an appreciable and/or statistically significant effect on any of the examined object features.

Keywords: Sensitivity analysis, statistical models, parametrization, NWP, analysis of variance.

1. Introduction

Stochastic Kinetic Energy Backscatter Schemes (SKEBS) are introduced in numerical weather forecast models to enhance their skill in the production of probabilistic forecasts. First introduced in Large Eddy Simulation models (e.g., Leith 1990; Mason and Thomson 1994), SKEBS are used to represent energetic contributions to flows from unresolved physical processes through stochastic perturbations. For atmospheric flows, these perturbations are added to model tendencies resulting in better calibrated forecast

ensembles (i.e., better match between mean errors and forecast uncertainties as represented by the variance in the ensemble). Such schemes are formulated, as with any other parameterization scheme, using a number of parameters that must be determined based on physical knowledge and intuition, or tuned to obtain a desired outcome such as increasing the variance in ensemble forecast members by a given amount. As such, it is of interest to understand what effect the SKEBS parameter values have on the evolution of simulated atmospheric states, especially if a specific effect is desired. For example, if increased ensemble variance is the end goal, then it is useful to know which parameters to vary to that end. Or, if one is performing object-oriented forecast verification (e.g., Gilleland *et al*., 2009; Marzban *et al*., 2009), then it is important to know how features of the objects are affected by model parameters. All of these issues can be examined under the umbrella of sensitivity analysis.

It is important to distinguish two distinct categories of Sensitivity Analysis (SA). In one category SA is done primarily for the purpose of model tuning and/or data assimilation, e.g., Ancell, Hakim (2007), Järvinen *et al.*, (2012), Laine *et al.*, (2012), and Ollinaho *et al.*, (2014). In this category the SA is only a component of a complex optimization problem where one seeks specific values of parameters (or initial conditions, etc.) that optimize some quantity gauging the agreement between forecasts and observations. Another way in which observations play a central role in this category of works is through data assimilation. By contrast, in the second category, SA does not involve any optimization or data assimilation (Alpert, 1993; Aires *et al.*, 2013; Marzban, 2013; Marzban *et al.*, 2014; Marzban *et al.* 2018a; Marzban *et al.* 2018b; Yang *et al.*, 2014; Dasari, Salgado, 2015; Smith, *et al.*, 2015); there the main purpose of SA is to assess the effect of the parameters on the forecasts. The main goal is not to optimize forecasts but rather gain knowledge on the relationship between model parameters and forecasts. This knowledge may, in turn, be used for improving forecasts, or it may shed light on the underlying physics of the phenomenon under study. There are (at least) two reasons that render this latter approach to SA nontrivial: 1) The effect, on forecasts, of a given parameter cannot be assessed independently of other parameters because the underlying physics is inherently multivariate, and 2) natural variability must be taken into account in order to establish the statistical significance of the results. Properly attending to these issues is a complicated task that has led to a large body of literature on this flavor of SA (Alpert 1993; Sobol', 1993; Oakley, O'Hagan, 2004; Fasso, 2006; Saltelli *et al.*, 2010; Zhao, Tiede, 2011; Aires *et al.*, 2013; Marzban, 2013; Marzban *et al.*, 2014; Marzban *et al*. 2018a; Marzban *et al.* 2018b). The present work falls into the latter category.

The approach adopted here consists of assessing the sensitivity of object features of meteorological interest. Four object types are considered: upper-air jet streaks, low level jets, precipitation areas, and frontal boundaries (i.e., baroclinic zones). Because the SKEBS parameters affect the amount of energy that is injected into the flow, one expects that large-scale features that rely on energetic growth (such as growing baroclinic modes) would be affected by different parameter values.

Here the SKEBS in the Weather Research and Forecasting (WRF) model (Skamarock, Klemp, 2008) is used. This SKEBS implementation introduces stochastic perturbations to the simulated tendencies of potential temperature and non-divergent wind, which are controlled through several user-specified parameters. Some of the parameters are deterministic in nature, such as those used to control the amplitude of the perturbations, which represent the total amount of backscattered energy in potential temperature and non-divergent wind. However, since the perturbations are generated using an Autoregressive process, there is also an element of pure randomness, hereafter referred to as the purely stochastic component of SKEBS. This component is controlled by a seed parameter that affects the random number generation in SKEBS. The reader is referred to Berner *et al.* (2011) for more details on this SKEBS implementation. Here, the effect of both types of parameters are evaluated and contrasted using WRF forecasts generated up to 120 hours. There exist many more parameters in SKEBS whose impact on features of objects is worthy of consideration. Here, the analysis is restricted to only two model parameters in order to simplify the demonstration of the methodology. (In a work to presented separately, as many as eight model parameters are being examined by the authors).

Serving as the central piece in this evaluation are the four aforementioned object types identified within gridded forecast fields. Section 4 describes a simple threshold-based method for identifying the objects. In addition to the number of identified objects, various quantities characterizing each object are recorded. For this study, these quantities serve as the response variable in linear models, and methods of experimental design provide a setting wherein the effect of several factors on these responses can be quantified.

Two of the factors are key SKEBS parameters (the amplitude of perturbations to rotational wind and potential temperature), and a third factor is the replication of SKEBS itself (i.e., the seed used to generate sequences of random perturbations in SKEBS); this factor represents the purely stochastic component of SKEBS. The fourth factor represents the effect of daily variability. The third factor can be viewed as generating an ensemble, and the fourth factor is motivated by the expectation that forecasts are sensitive to initial conditions. The effect of these four factors is estimated for forecast hours 0-120.

As further explained in Section 2, the design of the experiment involves nine days, 41 forecasts at 3hr intervals (between 0 and 120 hours), nine values of each of the two SKEBS parameters, and six SKEBS replications, which in a full factorial design leads to a large number of experiments (or “ensemble members”); to reduce the number of experiments, a special type of a fractional factorial design (called a Latin Square Design) is used.

Experiments of this type are often called *computer experiments* because the resulting data are not the result of a real experiment in any sense of the word (Sacks *et al.*, 1989; Welch *et al.*, 1992; Santner, *et al.*, 2003; Fang, *et al.*, 2006). The defining characteristic of computer experiments is that the experimental error is zero, because re-running the

computer model (here WRF/SKEBS) leads to the same set of outcomes. Without an estimate of experimental error it is impossible to perform any of the statistical tests designed to assess statistical significance (Santner *et al.*, 2003; Fang *et al.*, 2006). However, as long as one is interested in main effects only (i.e., no higher-order interactions), then standard methods of experimental design can be used for assessing statistical significance, because all of the contributions to variance from higher-order interactions can act as a proxy for experimental error (Montgomery, 2009).

2. Experimental Design: A Brief Introduction

This study aims to determine how certain spatial features of forecasts are affected by four factors, including two model parameters: the amplitude of perturbations to 1) rotational wind, and 2) potential temperature, denoted Par1 and Par2, respectively. Additionally, another factor is also examined – one that measures how the effects vary across (here, 9) days; it is denoted Day. One important question is: How does the effect of the deterministic parameters (Par1 and Par2) compare with the effect of the purely stochastic component of SKEBS? Therefore, in addition to the three factors Day, Par1, and Par2, a fourth factor - denoted Rep - is introduced to measure the effect of replicating the experiment. Finally, it is useful to examine how all these effects vary with forecast (valid) time, denoted Fhour (here, varying from 0 to 120 hours).

In the field of experimental design (Montgomery, 2009), linear models are often employed to estimate the effect of various factors on the response. One simple model is

$$y_{ijkl} = \mu + Day_i + Par1_j + Par2_k + Rep_l + \varepsilon_{ijkl} \quad , \qquad (1)$$

where the response y_{ijkl} denotes a measurement of some quantity of interest (e.g., the number of jet streaks) on the i^{th} Day, for the j^{th} and k^{th} values of Par1 and Par2, respectively, and for the l^{th} replication of the experiment. The factor Fhour is not included in the model, because the model is developed at each value of Fhour. The terms appearing on the right side of Eq. (1) are all parameters (not to be confused with SKEBS parameters) to be estimated from data on the response y and the factors. The ε term is a random variable whose variance σ_ε^2 is another quantity that must be estimated from data, not only for assessing goodness-of-fit, but also for performing statistical tests. It can be shown (Montgomery, 2009) that the least-squares estimates of these parameters generally involve sample means of the response, or the difference between two sample means. For example, the least-squares estimate of the μ parameter is the sample mean $y_{....}$, also called the *grand mean*. The parameter Day_1 is estimated by the difference $(y_{1...} - y_{....})$. In all of these expressions a "dot" refers to a sample mean over the corresponding index. The other components - the Day factor, and the other factors in the model - are all estimated through similar difference between sample means. Given that the estimates of the factors are differences from the grand mean, these estimates are also called *main effects*. The machinery of experimental design aims to perform statistical/hypothesis tests of whether the true/population main effects are zero; (see next paragraph for another measure of a factor's effect.) The model in Eq. (1) is strictly linear, but it is possible to introduce

nonlinear terms. Such terms generally appear as terms with multiple indices, and they are called *interaction effects*. For example, a term like X_{ij} (called a 2-way interaction) measures how the effect of Par1 on response varies across days.

Although tests of main effects are performed for the problem at hand, there exists an alternative approach which is also appropriate. Strictly speaking, the main effects discussed above are estimates of fixed, population parameters, and for this reason they are called *fixed effects*. Any conclusions based on a fixed effects model are specific only to the particular values assigned to the various factors. However, one may choose to view these particular values as a random sample taken from a larger space of parameter values, in which case it makes no sense to speak of the main effect of a factor, because any notion of an effect is itself a random variable. Effects of this type are called *random effects*, and any conclusions based on a random effects model pertain to the population of **all** possible values that the factors may take, not the specific values appearing in the sample only. In such models, the main aim is not to test whether or not an effect is zero, but rather to test whether or not any portion of the variability in the response can be explained by each of the factors in the model. Specifically, for random effects models one writes

$$\sigma^2_{Response} = \sigma^2_{Day} + \sigma^2_{Par1} + \sigma^2_{Par2} + \sigma^2_{Rep} + \sigma^2_{\varepsilon}, \qquad (2)$$

and the goal is to estimate and then test whether any of the *variance components* on the right hand side of Eq. (2) are zero.

To clarify the difference between a fixed effects model and a random effects model, suppose the Day factor takes d values (i.e., the number of days in the study). Treating the Day factor as a fixed factor would allow one to test whether there is a difference between the sample means of the response across the d days. A significant result would then suggest that the mean response varies across the specific d days, i.e., the Day factor has an effect on the response. However, one may choose to consider the d days in the study as a random sample taken from the population of all days, in which case it is more appropriate to treat the Day factor as a random factor. Then, one can test the null hypothesis $\sigma^2_{Day} = 0$ which constitutes a test of whether a nonzero portion of the total variability in the response $\sigma^2_{Response}$ can be accounted for by daily variability. A significant result would suggest that the mean response varies across **all** days (not just the d days appearing in the data). Similarly, one can treat Rep, Par1, and Par2 as fixed or random factors. Although fixed effects models provide intuitive measures of effects, random effects models have the advantage that the final conclusions are not specific to the values of the factors chosen for the study. As such, both model types are useful.

Therefore, here, both types of models are developed. First, the factors are treated as fixed parameters. The estimate of each factor represents the sensitivity of the response with respect to that factor, i.e., the main effect of that factor. Then, random effects models are developed wherein the sensitivity of the response with respect to a given factor is measured by the variance component of that factor. It is more useful to report the variance

component as the fraction of the total variance. For example, the sensitivity for the Day factor is best reported as the so-called *intraclass correlation*

$$\rho_{Day} = 100 \frac{\sigma^2_{DAY}}{\sigma^2_{Response}} \quad ; \qquad (3)$$

similarly for the other variance components. Another advantage of examining the intraclass correlation is that analytic formulas exist for its confidence intervals (Montgomery, 2009). Such confidence intervals are critical for assessing the statistical significance of the sensitivity results.

In a full factorial design involving the four factors Day, Par1, Par2, and Rep, the number of model runs would be equal to the product of the number of values of each factor. That number of runs is often impractically large, and so, there exist a number of experimental designs whose goal is to reduce the number of runs. The Latin Square Design (LSD) is one such design, and it is briefly explained in the Appendix. In order to illustrate the basic idea, consider a problem involving three factors (and a response), with each factor taking three possible values. Ideally, one must observe the response at all 3^3 possible values of the factors, because then one can estimate the effect of the three factors as well as all of the interactions between them. However, it can be shown (Montgomery, 2009) that 3^2 runs are sufficient for estimating the main effects of the factors, if the values of the 3 factors for the 3^2 runs are selected according to a special prescription best displayed as a square table. An example of such a square is shown in Table 1, where the factors are denoted A, B, and C, and the subscripts denote the value of each factor. For example, the bottom/right element in that square corresponds to a run where the factors A and B are set to their third value, and the factor C is set to its second value. If the three factors have p levels, then the square table is p × p, and so, the necessary number of runs is only p^2. This example involves three factors, but it can be shown that the number of necessary runs is p^2 regardless of the number of factors (See Appendix). Such tables are called Latin Squares, and by virtue of being square tables, designs that follow such tables dramatically reduce the number of necessary runs, although at the cost of making all interactions between the factors inestimable (Montgomery, 2009). The inability of the LSD to estimate interaction effects is not a major concern because the main effects are generally much larger than interaction effects. The expectation that higher-order interactions are weaker than main effects is generally borne out due to several principles: the principle of hierarchical ordering, the principle of effect sparsity, and the principle of effect hierarchy; see pages 192, 230, 272, 314, 329 in (Montgomery 2009), and pages 33-34 in (Li, Sudarsanam, and Frey 2006). In the case of precipitation, Marzban et al. (2014) also find the interactions to be much smaller than main effects.

3. Data

Version 3.7.0 of the WRF-ARW model was used for this work, with lateral boundary conditions specified every 6 hours from output of the Global Forecast System (GFS). All of the standard WRF parameters were the default "out of box" parameters, with a 25-*Km*

grid-spacing for a domain 200 (east-west) by 140 (north-south), over the Continental US. Nine days are selected between December 2014 and March 2015. Each initial forecast hour is 10 days apart in this time period, ensuring minimal temporal association between days. The specific dates are as follows: Dec. 01, 11, 21, 31, Jan. 10, 20, 30, and Feb. 9, and 19. Winter months were chosen for the high degree of variability with regards to jet stream activity and mid-latitude cyclone activity.

For this study three factors Day, Par1, and Par2 were sampled according to the LSD, thereby reducing the necessary number of runs from 9^3 to 9^2. As a result, it is assumed that the interactions between these three factors are much smaller than the main effects. Because of the LSD, Par1 and Par2 take nine values as well. The range of the nine values are chosen to be centered on the recommended SKEBS values, but in order to examine the full range of possible effects, they span one order of magnitude smaller and one order of magnitude larger than the default values. The nine specific values are (0.1, 0.325, 0.550, 0.775, 1.000 3.250 5.500, 7.750, 10.000) $\times 10^{-5}$ for Par1, and (0.1, 0.325, 0.550, 0.775, 1.000, 3.250, 5.500, 7.750, 10.000) $\times 10^{-6}$ for Par2. As mentioned previously, in the random effects model inference of the sensitivities pertains to all possible values of the parameters, not just to the specific nine values; for this reason, the specific nine values selected here do not play an important role in the final analysis. Indeed, in an earlier version of the analysis, the following Par1 values produced very similar results: (0.5, 2.875, 5.25, 7.625, 1.0, 12.375, 14.75, 17.125, 19.5) $\times$ 10^{-5}.

One of the main goals here is to assess the effect of Rep (i.e., the purely stochastic component of SKEBS) and how it compares with the effect of the other factors. Therefore, more computational effort is dedicated to that end. Specifically, the Rep factor and the combination of the other three factors (Day, Par1, and Par2) follow a full factorial design. In other words, all 9^2 LSD runs involving Day, Par1, and Par2, are replicated at every value of the Rep factor. Here the number of replications is six.

All of these factors are treated first as fixed factors, and then as random factors. However, the factor Fhour is treated as a fixed factor, because it varies across 41 fixed values, from 0 to 120 hours, in increments of 3 hours. As mentioned above, Fhour is not included in the model, because the model in Eq. (1) is developed at each of the 41 values of Fhour. Consequently, all of the results found here take the form of "time series" of the main effects, variance components, or intraclass correlations as a function of Fhour.

Given the above design, the total number of runs is $9^2 \times 6 \times 41 = 19{,}926$. Although this is a large number of runs, it is significantly smaller than what would be necessary in a full factorial design: $9^3 \times 6 \times 41 = 179{,}334$.

4. The Response

As previously mentioned, sensitivity to SKEBS parameters is assessed with respect to various features of the following meteorologically significant object types: 1) upper-air jet streaks, 2) low-level jets, 3) precipitation events , and 4) frontal boundaries. Jet streaks

are defined at 250 *hPa* as regions with contiguous model grid points having wind speeds in excess of 50 $m\ s^{-1}$ (approximately 100 *knots*). Similarly, low level jets are defined as regions at 850 *hPa* characterized by winds stronger than 20 $m\ s^{-1}$. Precipitation events are contiguous regions where the total precipitation accumulation at the surface is above 1*mm* in 3-hour forecast intervals. Frontal boundaries (i.e., baroclinic zones) are identified using the horizontal gradients of the 1000-700 *hPa* geopotential thickness field (McCann, Whistler, 2001). A threshold value of 0.3 $m\ Km^{-1}$ for the magnitude of the geopotential height gradient is used to identify significant baroclinic zones. Although it is possible to develop more sophisticated means of identifying such objects, the focus of this study is on the development of an object-based SA method, regardless of how the objects are identified.

By definition, all of these objects are characterized by relatively well-defined spatial extent. For every available forecast hour, objects meeting the above criteria are identified. Figure 1 shows an example of jet streak objects identified in the 250 *hPa* WRF wind field. Three jet streak objects are identified in this particular WRF forecast. The smallest and weakest is located over states in the northwestern United States, a second is located over eastern Canada, while the largest and strongest extends from the southwestern Four Corners states to the mid-Atlantic states.

Here we point out that the five grid points nearest the model domain's lateral boundaries are omitted from the analysis in order to prevent any direct influence from the imposed lateral boundary conditions taken from deterministic GFS forecasts. This way, only grid points in the interior of the domain where solutions are fully influenced by SKEBS perturbations are considered. Also, identified objects are restricted to those composed of at least 50 grid cells, corresponding to areas larger than about 31,000 Km^2 in order to minimize any "noise" in the resulting object datasets that could be associated with spurious appearance/disappearance of small areas with wind speeds changing to values just above/below the threshold. Despite the application of such conditions, spurious changes in object characteristics may occur as objects merge or separate solely due to subtle changes in the underlying continuous field; for example, it is possible for two nearby jet streak objects at a particular forecast hour to merge at the following forecast hour due to an increase in wind speed above the threshold in the region separating the two jets. Associated changes to the response variable (e.g., number of objects or their size, intensity, and location) can be described as "measurement error" because the variability introduced by these changes is not due to any of the factors included in the model (Eq. 1).

The features examined here are 1) the number of objects, 2) their size, 3) intensity, and 4) location. The size of each object is computed as the number of grid points included in that object. The intensity is measured as the mean intensity of the field across the object, and their location is recorded at the latitude and longitude of the center-of-mass of the object. Panels a-c in Figure 2 show the histograms of number, size, and intensity for precipitation objects across all four factors (i.e., days, replications, Par1 and Par2); other object types have similar histograms. The histograms of latitude and longitude are not shown because

that figure shows no useful information. It can be seen that the number of objects can vary between 1 and 13, with the most common value around 3 or 4. By contrast, the size of objects has an exponential-looking histogram, and so the data examined consists of mostly small objects (i.e., consisting of 50 grid points). Mean intensity values (panel c) vary between about 1 and 17 *m/s*, with the most common value around 2.3 *m/s*.

Given the similarity in the shape of the histograms of number and size of objects, one may wonder if these two features are correlated. In fact, given that the size of the forecast domain is fixed, one may suspect a negative correlation. Panel d in Figure 2 shows the scatterplot of these two features. Although for the extreme case where there are as many as 13 objects, their size is restricted to be around 500 grid points, for cases with four objects, their size can vary from the smallest possible value (50) to 3500 grid points. As such, it can be seen that there is no linear association between the two features.

The histograms discussed above are constructed from the object features that arise in the data across all values of the four factors. But even for given values of the four factors, there exists a distribution of features. Here, that distribution is summarized by two quantities - the minimum and maximum; (the 25^{th} and 75^{th} percentiles of the histograms were also examined, but the results were statistically equivalent to those based on the minimum and maximum) In short, the aim is to study the effect of the aforementioned four factors on the following response/feature variables: 1) Number of objects (e.g., jet streaks) across the forecast domain, 2) Minimum and 3) Maximum size of (i.e., smallest and largest) objects across the domain, and 4) Minimum and 5) Maximum intensity (i.e., weakest and strongest) of objects across the domain. As for the location feature, the minimum, median, and maximum of latitude and longitude are also examined; it can be argued that the two SKEBS parameters considered here may have an effect on the location of the objects because they control propagation and development rates. (The authors acknowledge an anonymous Reviewer for this suggestion).

Although all four object types (upper-air jet streaks, low-level jets, precipitation events, and frontal boundaries) have been analyzed, only sensitivity results pertaining to upper-air jet streaks are shown in the next section. Results with respect to the other object types were found to be similar especially in terms of the relative magnitude of the effect of the four factors. Of the various features considered here, specific results pertaining to latitude and longitude are not shown, because they are similar to those pertaining to the intensity feature.

5. Results

Before developing the aforementioned models, it is useful to examine the simulated data, first. Figure 3 shows the values of the five responses/features as a function of forecast time (Fhour), on one day, with model parameters set to default values, and for the six replications (in colours). The thick/black line corresponds to a run wherein all of SKEBS has been turned off. It can be seen that the coloured curves (i.e., different replications of SKEBS with default parameters) generally fluctuate about the curve of this control run.

Moreover, evidently, all five response variables have significant variability across forecast times. Part of this variability is "real" in the sense that objects can appear and disappear in a forecast field across three hours. The remainder of the variability is due to the aforementioned measurement error; for example, although the actual size of an object may not change in a 3-hour interval, the thresholding procedure adopted here for identifying objects may give a slightly different value for the size. This measurement error is not a stumbling block for the analysis; its only effect is to magnify the variance of the ε term in Eq. (1), and thereby reduce statistical power. Also, as mentioned at the end of Section 1, these empirical errors are necessary for performing statistical tests of significance in computer experiments.

The variability of the response variable plays an important role in both fixed effects and random effects models. Figure 4 shows the variability of the five response variables at each forecast time. The slow modulations of all of these curves correspond to the natural evolution of weather patterns in the nine days examined here. To obtain a sense of the variability of these results, 95% confidence intervals are also shown (as vertical bars). It is evident that all five response variables have nonzero variance at all forecast hours. Recall that the goal of random effects models is to determine how these variances are apportioned across the various factors in the model.

The linear model in Eq. (1) is developed at each forecast time. Treating the factors as fixed factors allows one to perform F (or t) tests on the main effects. The resulting p-values are summarized in Figure 5. The variability in the boxplots is across the 120 forecast hours. Here, a significance level (e.g., 0.05 or 0.01) is not selected to assess statistical significance. Instead, the boxplot of the p-values is examined to provide a visual assessment of the "strength" of the statistical significance. A tight boxplot, near 0, suggests that the corresponding effect is statistically significant. By contrast, if the boxplot of p-values is near 1 or extends across the full range from 0 to 1, then the corresponding factor is deemed non-significant, i.e., that there is insufficient evidence from data to conclude that the factor has an effect. This practice is consistent with a fundamental theorem in statistics stating that the distribution of p-values is given by a uniform distribution between 0 and 1, if the null hypothesis (of no-effect) is true.

Here (Figure 5) it can be seen that the factor Day has a significant effect on all five responses. This is not surprising, because it is known that the responses vary across the nine days in the study. By contrast, the near-1 location of the boxplots for Par1 and Par2 in all five panels suggests that there is no evidence from data to suggest that these two parameters have any effect on any of the response variables. The Rep factor plays a more complex role; although the p-values do extend to relatively large values, the bulk of their histogram is skewed toward smaller values, in all five panels. In other words, the Rep factor does appear to have an effect on all five response variables, but not at all forecast hours.

Although it is possible to examine the p-values in the fixed-effects model at each forecast hour, it is more useful to examine the forecast-hour-dependence of the results in the

random effects model. Treating the factors as random variables leads to consideration of the variance components, and in turn, intraclass correlations, ρ in Eq. (3), and their confidence intervals. Figure 6 shows the 95% confidence interval for ρ at different forecast times. The "Day, Number" panel shows the effect of the Day factor on the number of objects in the domain. It can be seen that the effect of the Day factor diminishes very quickly, and falls to near-zero values for Fhour beyond nine. The effect of the Day factor on the size of objects is shown in the panel marked "Day, Size"; although the smallest (black) objects are mostly unaffected by the Day factor, the effect on the largest (red) objects is less trivial. On forecast times scales from 0 to 120 hours, for very short forecast times (3 to 9 hours) the Day factor can explain 60% to 90% of the variability; even for longer forecasts, the effect is non-zero, leveling-off at values in the 5% to 10% range. In other words, even for very long forecast times, daily variability contributes a significant portion of the total variability in the size of objects. The effect of the Day factor on the (mean) intensity of objects has a similar behavior (panel "Day, Intensity"), although for longer forecast times, the effect is generally weaker than the effect on object size. Said differently, for short forecast hours the variability in object intensity can be explained by daily variability, but for very long forecast times that variability is not due to daily changes in weather.

The effect of the Rep factor can be seen in the second row of panels in Figure 6. For all five response variables (number, min. size, max. size, min. intensity, and max. intensity), Rep can explain only about 0.1% to 0.5% of the variability in the data. The large confidence intervals make it difficult to interpret the results; the lower end of the intervals are generally above zero, suggesting that the corresponding effects are nonzero, consistent with the small p-values observed in Figure 5. Although the top end of the intervals is erratic, it is important to note the scale on the y-axis of these panels - 0 to 1% - and so, the effect of Rep is generally quite small.

The effect of the parameters (Par1, Par2) on all response variables is even weaker than that of the Rep factor (third and fourth rows in Figure 6). The ρ values are generally below 0.1%. In other words, even when the effect of the parameters is statistically significant (i.e., nonzero at 95% confidence level), the magnitude of the effects is extremely small. The fact that the effect of the parameters is weaker than that of Rep is important, and is further discussed in the next section.

The last row of panels in Figure 6 shows ρ_ε, i.e., the percentage of the variability in the data that cannot be explained by the four factors Day, Rep, Par1, and Par2. As such, it is useful for assessing the combined effect of the four factors. Evidently, for forecast hours longer than three hours nearly 100% of the variability in the number of objects cannot be explained by any of the four factors. This is expected from the panels in the first column of Figure 6, because none of the four factors appear to have an effect on the number of objects for long forecast times.

When the response is object size (bottom row, middle panel), or object intensity (bottom row, right panel), the variability that cannot be explained by the four factors generally

increases with forecast time. For the smallest of objects (black curve) the increase is quite abrupt – from 0 to 100% as one goes from 0hr to 3hr forecasts and beyond. For the largest objects (red curve), although the increase is more gradual, the percentage of unexplained variance approaches 100% by forecast hour 100. The undulations in the curves, caused by the natural variability in the data across the 120 hours, make it difficult to pinpoint a specific forecast time beyond which the four factors become useless.

In summary, examining all of the panels in Figure 6, it appears that when the factors do contribute to the variability in the response, most of that variability is due to the Day factor. The next important factor is Rep; and Par1 and Par2 have nearly no effect. It is also clear that Par1 and Par2 have a much smaller effect than Rep, at every forecast hour. This suggests that the two tunable SKEBS parameters examined here may not produce the expected variability in the specific objects under consideration, since the purely stochastic component (which is not as controllable as the tunable parameters) overwhelms the variability in the forecasts.

6. Conclusion and Discussion

SKEBS has been designed to introduce variability into the forecasts in a manner consistent with the physics that are unresolved by the model. One would then expect that SKEBS parameters (Par1, Par2) would have some effect on the forecasts, and that the effect of these parameters would be more prominent than that of the purely stochastic component of SKEBS (Rep). Here, forecasts of jet streaks, low-level jets, precipitation, and baroclinic zones are considered, although only the analysis on jet streaks is presented. A simple method is employed to identify these objects within continuous forecast fields; a suite of methods from experimental design are then woven together to assess the effect of four factors (Day, Par1, Par2, Rep) on five features of these objects (number, minimum and maximum size, and minimum and maximum intensity. The impact of the four factors on the location (latitude and longitude) of the objects is also examined; but it is not presented because the results are similar to that of intensity. It is shown that the number of objects in these fields does not appear to be affected by any of the factors. It is also shown that for forecast times when the factors do have a nonzero effect on the size and intensity of objects, apart from the effect of the Day factor, the effects of the other three factors are quite small, explaining only a few percentage points of the variability observed in the data. More importantly, it is found that the effect of Par1 and Par2 is much less than that of Rep.

This suggests that the variability produced by varying the two SKEBS parameters does not appear to have a significant effect on the specific object types and their features examined here; the purely stochastic part is the main driver of any SKEBS-induced variability. It is important to emphasise that this conclusion pertains only to the specific object types and features examined here. It does not reflect on the connection between SKEBS and the physical processes it seeks to represent, and whether the physically-motivated model behind SKEBS has a consistent effect on model forecast evolution at large. In practice, then, if one is interested in the specific objects and features examined

here, it is best if the resources for tuning or calibrating the model parameters are directed away from the physical SKEBS parameters. However, see next paragraph.

Armed with the methodology developed here, the above analysis can be generalized in a number of ways. For instance, the criteria for identifying objects can be revised; the number of parameters, and their range and values can be extended, and/or other response variables can be examined. Although the two SKEBS parameters under consideration do not appear to have an effect on the four object types examined here, it will be useful to find other meteorologically relevant objects that are affected by these SKEBS parameters. As pointed out by an anonymous reviewer, it is known that the SKEBS parameters examined here do affect the reliability/skill of large-scale ensemble forecasts. As such, the null effect of the model parameters may seem contradictory, but then it is important to recall that the sizes of the objects considered here fall on the smaller end of the resolved scales in the model simulations.

One may also consider more/other SKEBS parameters, in which case Graeco-Latin Square Designs (GLSD; see Appendix) can be used to reduce the number of runs necessary for estimating main effects. A desirable feature of GLSD is that the necessary number of runs for estimating main effects is the square of the number of values each factor takes, independent of the number of factors in the study. In fact, fixed-effects and random-effects models with as many as eight parameters are currently under investigation, and preliminary results suggest that even when some of the SKEBS parameters do affect the spatial structure of the forecasts, their effect is still overwhelmed by daily variability and variability due to replication. That finding also raises the possibility of examining the effect of the factors on the spatial structure of the forecasts, independently of the existence of any objects in the forecast field. Generalizations can also be made to the statistical modeling effort. For example, the fixed-effects and random-effects models employed here are linear models commonly employed in experimental design (Montgomery 2009). These can be generalized to include higher order interactions. Alternatively, it is possible to replace these models entirely with fully nonlinear models - often called metamodels (Santner *et al.*, 2003; Aires, 2013). Many of these questions are currently under consideration.

A comparison of the current work and that reported in Marzban *et al.* 2018b is in order. First, and foremost, whereas the objects here are identified by a simple thresholding method, those in the latter work are identified via two different clustering algorithms. Second, the (11) model parameters in the latter study are continuous parameters which necessitates a different (than LSD) method for sampling the parameter space. The reason the model parameters are different between the two studies is that the underlying model in the latter work is COAMPS® (Coupled Ocean/Atmosphere Mesoscale Prediction System). The impact of the 11 parameters in COAMPS on the spatial structure of forecasts (i.e., without reference to any objects) has also been examined (Marzban *et al.* 2018a).

7. Appendix: Latin Square Designs

Consider an experiment involving three factors, *A, B,* and *C*, each taking three values denoted *A1, A2, A3, B1, B2, B3*, and *C1, C2, C3*. (In statistics, the values a discrete variable can take are referred to as levels. Here we avoid the term level in order to minimize confusion with the use of that term in meteorology.) A full factorial design refers to 3^3 runs necessary to consider all possible combinations of the values each factor can take. It can be shown (Montgomery, 2009) that in a full factorial design one can estimate all main effects, all interactions, and the variance of the errors, σ^2_ε . If, however, interactions are not of interest, then only the specific runs shown in Table 1 are sufficient. In other words, only the nine runs *(A1, B1, C1), (A1, B2, C2), (A1, B3, C3), ... , (A3, B3, C3)* are sufficient for estimating the main effects (and the error variance). An experiment involving only such specific runs is said to follow a Latin Square Design (LSD). Interactions, however, cannot be estimated. Technically, in LSD, main effects and interactions effects are said to be *aliased,* meaning that the effects one can estimate are a combination of main effects and interaction effects, and one cannot disentangle the two. As such, when one computes main effects in an LSD, the assumption is that the interaction effects are negligible. Latin squares as in Table 1 are constructed by assigning the columns to the values of one factor, the rows to the values of another factor, and then cyclicly permuting the values of the last factor within the body of the square. This assures that every combination of the three values appears precisely one time - a unique and defining characteristic of the LSD. The factors may take more than three values, in which case the Latin square will simply be larger.

The Graeco-Latin Square Design (GLSD) is the generalization of the LSD to four or more factors, with each factor taking any number of values; the only constraint is that all factors must have the same number of values. So, in the present study, if nine days are selected for the analysis, then each of the two parameters (Par1 and Par2) must take nine values. More examples of LSDs and GLSDs can be found in Montgomery (2009).

It is worth mentioning that in an LSD involving the three factors Day, Par1 and Par2, on no single day are the two parameters varied across all their values. Consequently, one cannot assess the sensitivity of the two parameters for each day. This may appear to be a limitation; however, it is important to point out that knowledge of sensitivities for any given day is useless; only the sensitivities across all days have practical utility. And the LSD allows one to estimate those sensitivities with only 9^2 runs (instead of 9^3).

It is important to distinguish LSDs (or GLSDs) with another sampling design with a similar name, namely Latin Hypercube Sampling (LHS). Although frequently used in SA (Hacker *et al.*, 2011; Marzban, 2013; Marzban *et al.*, 2014), the LHS is a completely different sampling scheme, and is most suitable for situations where the covariates (independent variables) are continuous quantities, not discrete factors; there, one specifies the desired sample size, first. Then, each of the factors is subdivided into that many bins, and a sample is drawn such that any combination of the bins appears precisely one time. The utility of the LHS derives from the fact that LHS estimates of model parameters are more precise (at least, no less precise) than estimates based on Simple Random Sampling

(McKay *et al.*, 1979). Note that by contrast to the LSD (or GLSD) where the sample size is simply the square of the number of values in a factor, the sample size in LHS is not determined by the number of values of a factor, or the number of factors; instead, it is specified by the user.

8. Acknowledgments

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

9. References

Aires F, Gentine P, Findell K, Lintner B, Kerr C. 2013. Neural Network-Based Sensitivity Analysis of Summertime Convection over the Continental United States. *J. Climate*, **27,** 1958-1979.

Ancell B, and Hakim G. 2007. Comparing Adjoint- and Ensemble-Sensitivity Analysis with Applications to Observation Targeting. *Mon. Weather. Rev.*, **135**, 4117-4134.

Berner J, Ha SY, Hacker JP, Fournier A, Snyder C. 2011. Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multiphysics representations. *Mon. Weather. Rev*., **139**, 1972-1995.

Dasari HP, Salgado R. 2015. Numerical modeling of heavy rainfall event over Madeira Island in Portugal: sensitivity to different micro physical processes. *Meteorol. Appl.*, **22**, 113-127.

Gilleland E, Ahijevych D, Brown BG, Casati B, Ebert E. 2009. Inter comparison of spatial forecast verification methods. Wes. *Forecasting*, **24**, 1416-1430.

Fasso A. 2006. Sensitivity Analysis for Environmental Models and Monitoring Networks. In: Voinov A, Jakeman AJ, Rizzoli, AE (eds). Proceedings of the iEMSs Third Biennial Meeting: Summit on Environmental Modeling and Software. International Environmental Modeling and Software Society, Burlington, USA, July 2006.
Internet: http://www.iemss.org/iemss2006/sessions/all.html

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

Hacker JP, Snyder C, Ha S-Y, Pocernich M. 2011. Linear and non-linear response to parameter variations in a mesoscale model. *Tellus A*, **63**, 429–444.

Järvinen H, Laine M, Solonen A, Haario H. 2012. Ensemble prediction and parameter estimation system: the concept. *Q. J. R. Meteorol. Soc.*, **138**, 281-288.

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

Leith CE. 1990. Stochastic backscatter in a subgrid-scale model: Plane shear mixing layer. *Physics of Fluids A: Fluid Dynamics*, **2.3**, 297-299.

Ollinaho P, Järvinen H., Bauer P, Laine M, Bechtold P, Susiluoto J, Haario H. 2014. Optimization of NWP model closure parameters using total energy norm of forecast error as a target. *Geoscientific Model Development*, **7**, 1889-1900.

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

Marzban C. 2013. Variance-based Sensitivity analysis: An illustration on the Lorenz '63 model. *Mon. Weather. Rev.*, **141**, 4069-4079.

Marzban C, Sandgathe S, Doyle JD, Lederer NC. 2014. Variance-based sensitivity analysis: Preliminary results in COAMPS. *Mon. Weather. Rev.*, **142**, 2028-2042.

Marzban, C, Du X, Sandgathe S, Doyle JD, Jin Y, Lederer NC. 2018a: Sensitivity analysis of the spatial structure of forecasts in mesoscale models: Continuous model parameters. *Mon. Wea. Rev.* **146**, 967-983.

Marzban, C, Jones C, Li N, Sandgathe S. 2018b: On the effect of model parameters on forecast objects. *Geoscientific Model Developmet,* **11**, 1-14.

Mason PJ, Thomson DJ. 1994. Stochastic backscatter in large-eddy simulations of boundary layers. *Journal of Fluid Mechanics*, **242**, 51-78.

McCann DW, Whistler JP. 2001. Problems and solutions for drawing fronts objectively. *Meteorol. Appl.*, **8**, 195-203.

McKay MD. Beckman RJ, Conover WJ. 1979. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics*, **21**, 239-245 .

Montgomery DC. 2009. *Design and Analysis of Experiments*, 7th Edition, John Wiley & Sons, 656 pp.

Oakley JE, O'Hagan A. 2004. Probabilistic sensitivity analysis of complex models: a Bayesian approach. *J. R. Statist. Soc., B,* **66**, 751-769.

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

Sacks J, Welch WJ, Mitchell TJ, Wynn HP. 1989. Design and Analysis of Computer Experiments. *Statistical Science*, **4**, 409-423.

Santner TJ, Williams BJ, Notz WI. 2003. *The Design and Analysis of Computer Experiments.* Springer, 299pp.

Smith SA, Vosper SB, Field PR. 2015. Sensitivity of orographic precipitation enhancement to horizontal resolution in the operational Met Office Weather forecasts. *Meterolog. Appl*., **22**, 14-24.

Skamarock WC, Klemp JB. 2008. A time-split nonhydrostatic atmospheric model for weather research and forecasting applications. *J. Comp. Phys.*, **227**, 3465-3485.

Sobol' IM. 1993. Sensitivity estimates for nonlinear mathematical models. *Mathematical Modeling and Computational Experiments*, **1,** 407-414.

Stein U. Alpert P. 1993. Factor separation in numerical simulations. *J. Atmos. Sci.*, **50**, 2107-2115.

Welch WJ, Buck RJ, Sacks J, Wynn HP, Mitchell TJ, Morris MD. 1992. Screening, Predicting, and Computer Experiments. *Technometrics*, **34**, 15-25.

Yang Y. Uddstrom M, Revell M, Moore S. 2014. Soil moisture simulation by JULES in New Zealand: verification and sensitivity tests. *Meteorol. Appl.*, **21**, 888-897.

Zhao J, Tiede C. 2011. Using a variance-based sensitivity analysis for analyzing the relation between measurements and unknown parameters of a physical model. *Nonlin. Processes Geophys.*, **18**, 269276.

Figure Captions

Figure 1. The histogram of a) the number of precipitation objects, and their b) size and c) intensity. Panel d shows the scatterplot of size versus the number of objects.

Figure 2. Jet streak objects identified within the WRF wind field at 250 hPa for a 42-hour forecast initialized at 00 UTC on February 9 2015. Jet streaks are identified by white contour lines, and the location of the maximum wind speed within each object is identified by the white-contoured black dot.

Figure 3. The "time series" of the five response variables: The Number of objects (a), the size of the smallest (b) and largest (c) objects, and the intensity of the weakest (d) and strongest (e) objects. The colors correspond to the six replications, and the thick/black line corresponds to a control run wherein SKEBS has been turned off. Par1 and Par2 are set to their default SKEBS values (10^{-5} and 10^{-6}, respectively). Size refers to the number of grid points in an object, and intensity is measured in m/s.

Figure 4. The variance (across all factors - Day, Rep, Par1, and Par2) of the five response variables - Number of objects (top), minimum size (black) and maximum size (red) (middle panel), and minimum intensity (black) and maximum intensity (red) (bottom panel). The vertical lines are 95% confidence intervals, displaying the uncertainty in these variance estimates. Size refers to the number of grid points in an object, and intensity is measured in m/s.

Figure 5. The distribution/boxplot (across 120 forecast hours) of p-values testing the significance of the main effects for the four factors (Day, Rep, Par1, Par2) on the five responses (panels a-e).

Figure 6. The 95% confidence intervals for the intraclass correlation ρ versus forecast time, displaying the effect of the four factors, Day, Rep, Par1 and Par2 (top 4 rows) on the five responses - number of objects (left column), minimum (black) and maximum (red) size (middle column), and minimum (black) and maximum (red) intensity (right column). The last row shows ρ_ε, the proportion of total variance in the response not explained by the four factors.

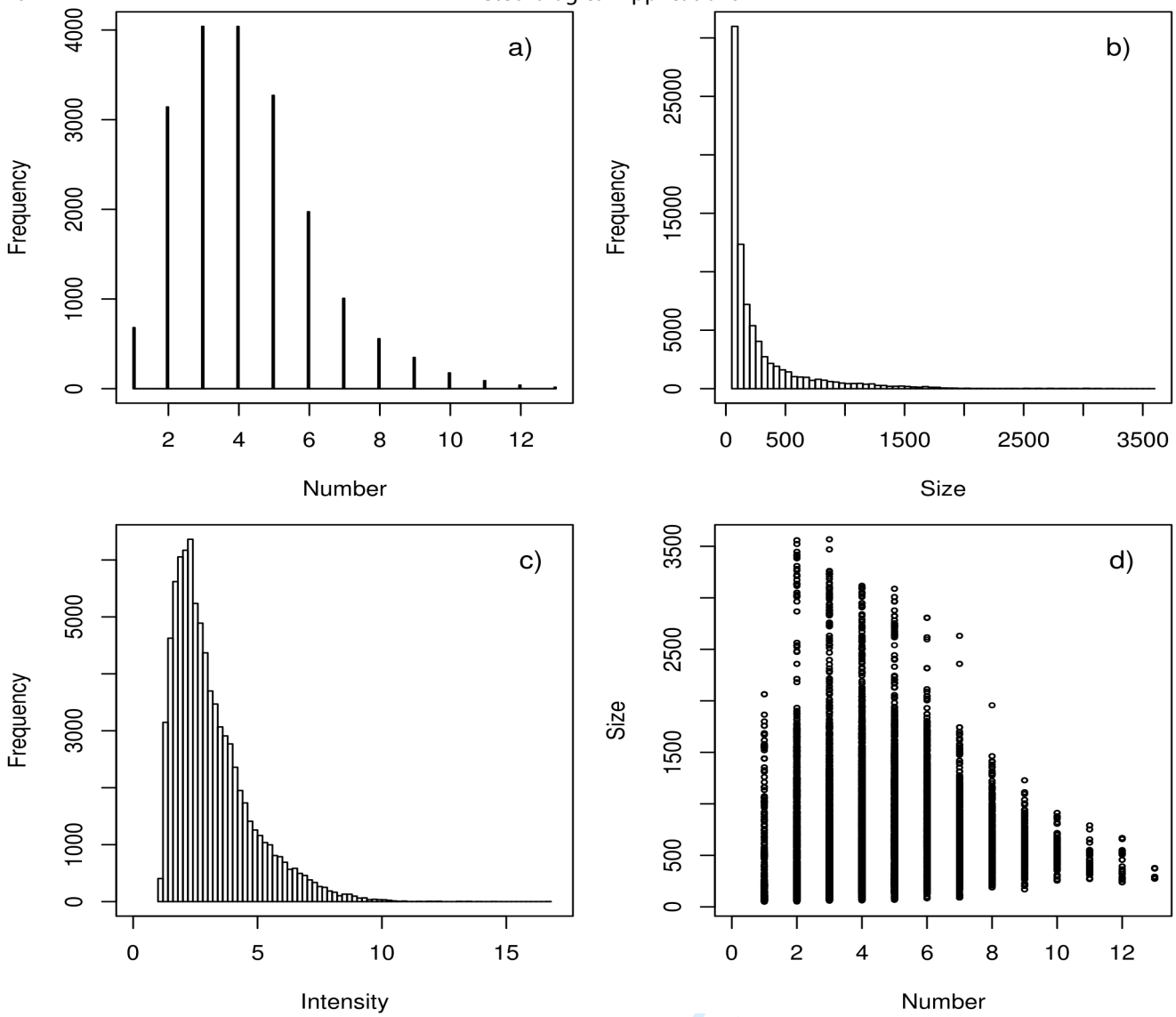

Figure 1. The histogram of a) the number of precipitation objects, and their b) size and c) intensity. Panel d shows the scatterplot of size versus the number of objects.

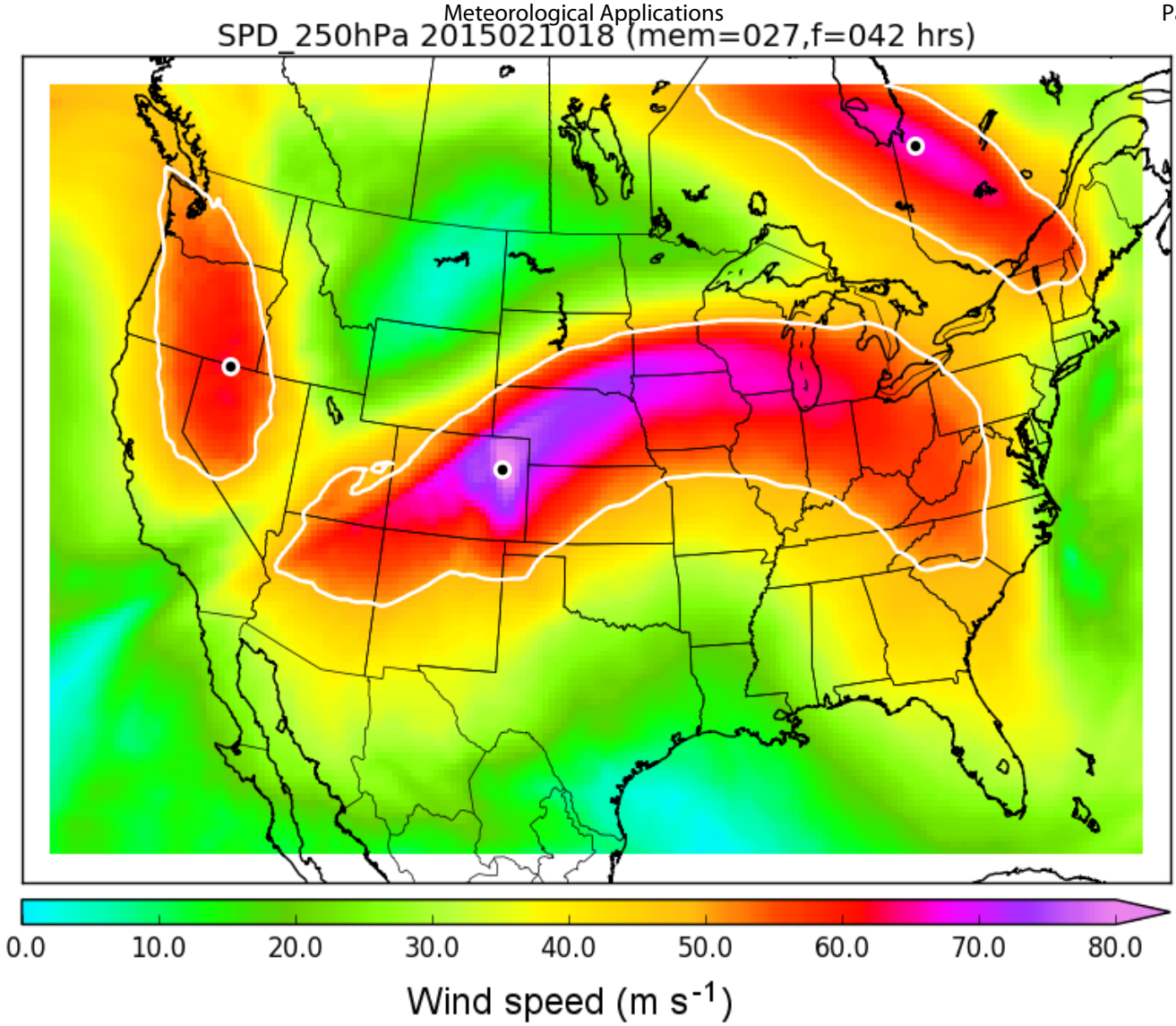

Figure 2. Jet streak objects identified within the WRF wind field at 250 hPa for a 42-hour forecast initialized at 00 UTC on February 9 2015. Jet streaks are identified by white contour lines, and the location of the maximum wind speed within each object is identified by the white-contoured black dot.

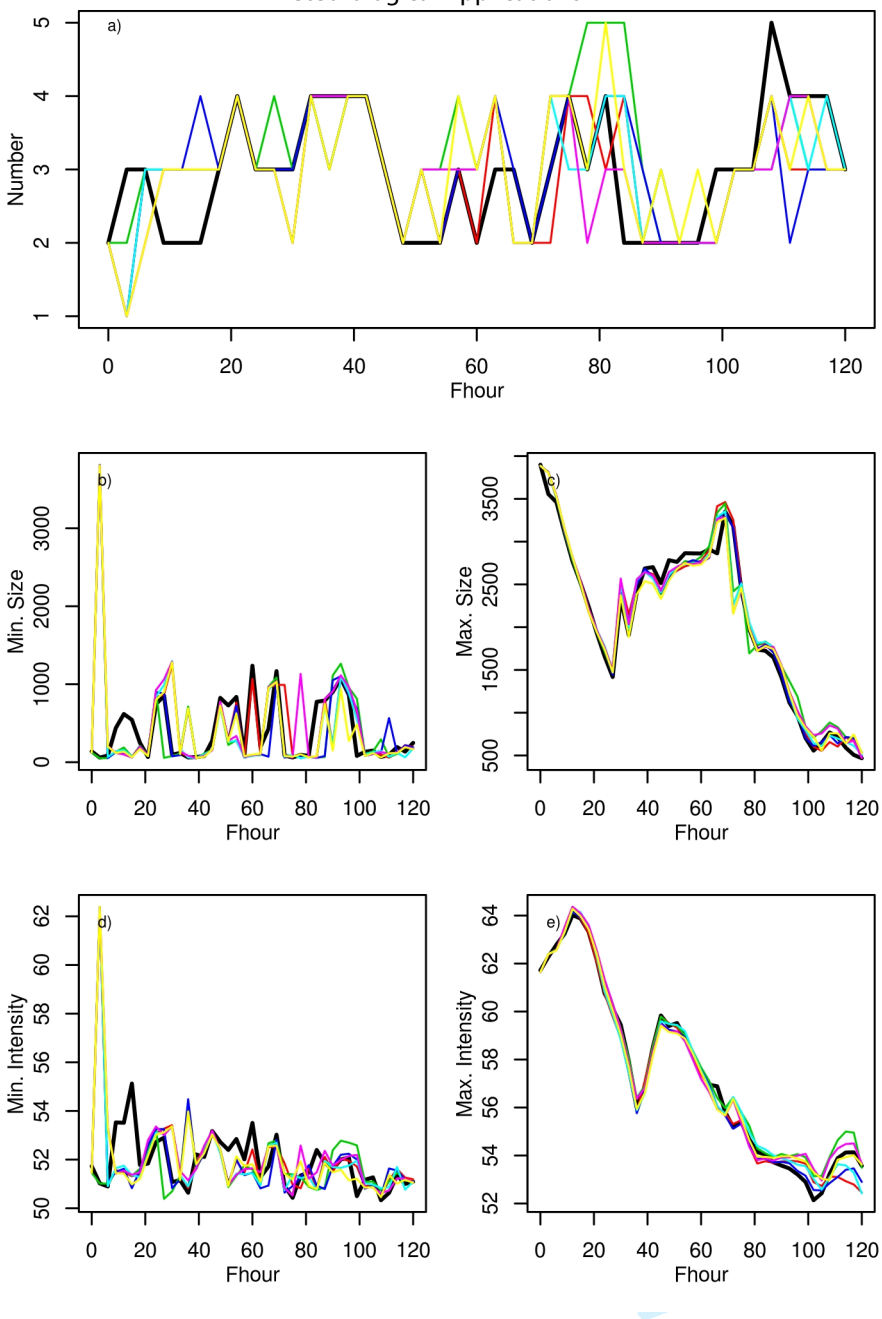

Figure 3. The "time series of the five response variables: The Number of objects (a), the size of the smallest (b) and largest (c) objects, and the intensity of the weakest (d) and strongest (e) objects. The colors correspond to the six replications, and the thick/black line corresponds to a control run wherein SKEBS has been turned off. Par1 and Par2 are set to their default SKEBS values (10^{-5} and 10^{-6}, respectively). Size refers to the number of grid points in an object, and intensity is measured in m/s.

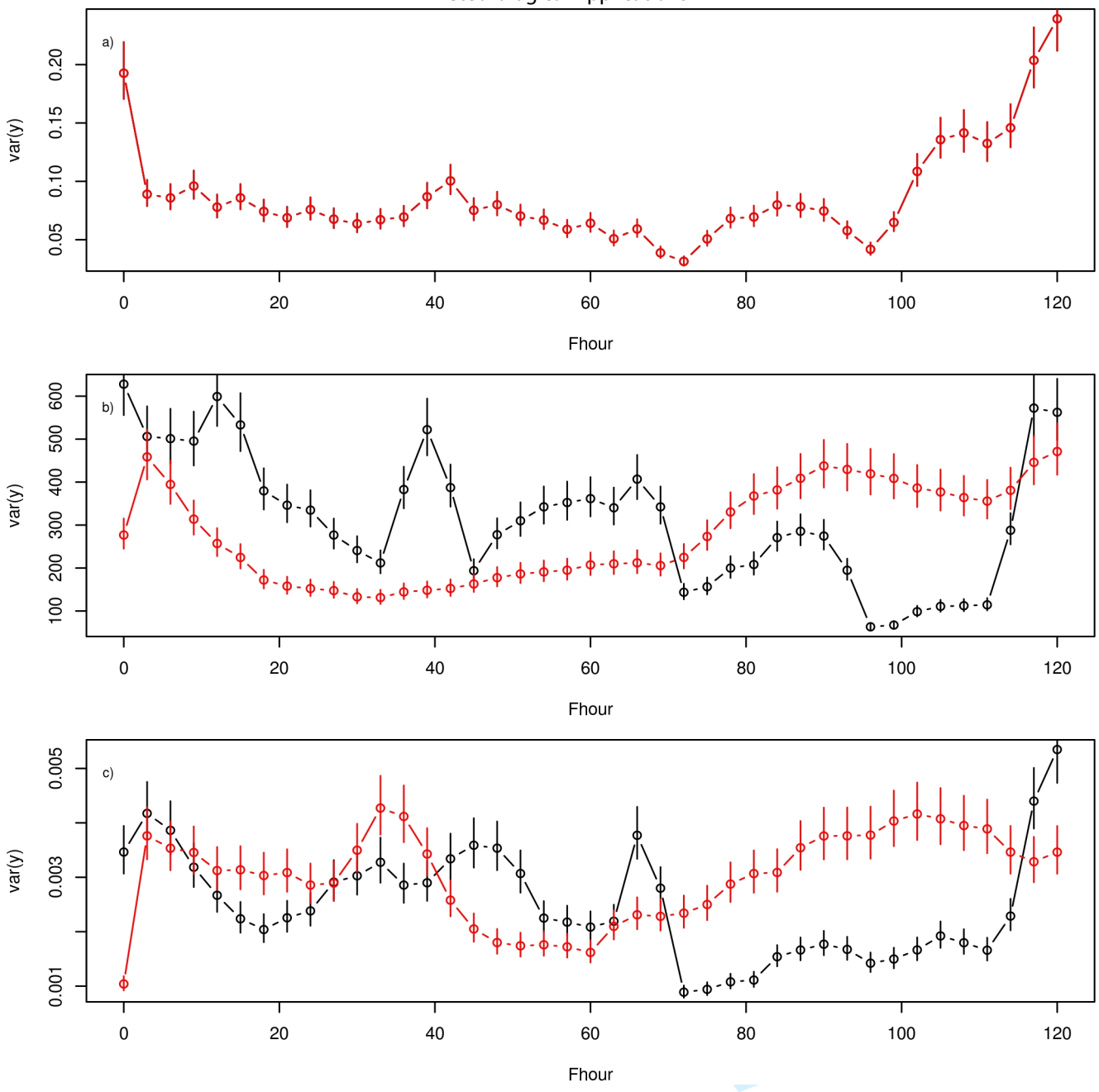

Figure 4. The variance (across all factors - Day, Rep, Par1, and Par2) of the five response variables - Number of objects (top), minimum size (black) and maximum size (red) (middle panel), and minimum intensity (black) and maximum intensity (red) (bottom panel). The vertical lines are 95% confidence intervals, displaying a sense of the uncertainty in these variance estimates. Size refers to the number of grid points in an object, and intensity is measured in m/s.

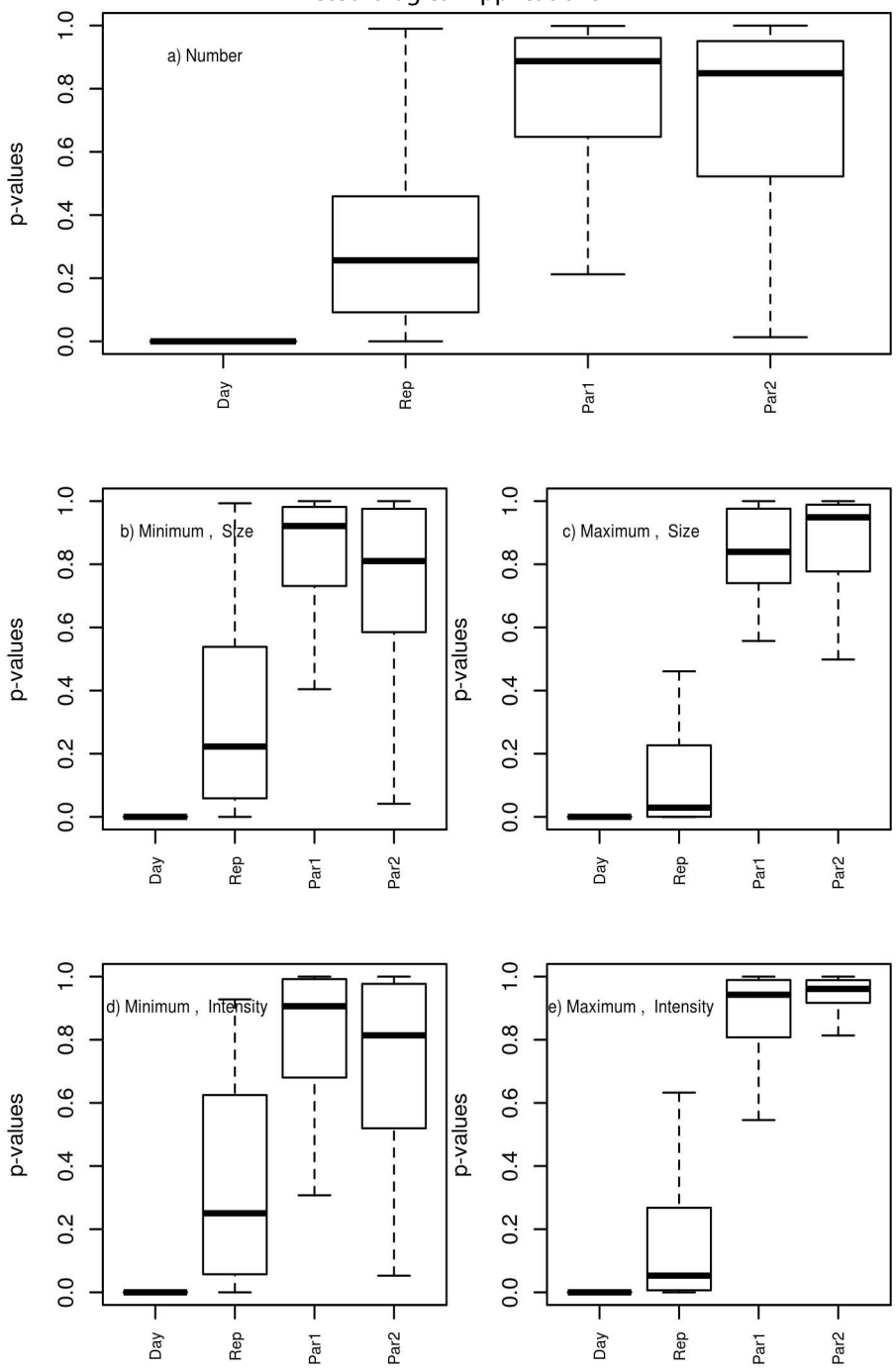

Figure 5. The distribution/boxplot (across 120 forecast hours) of p-values testing the significance of the main effects for the four factors (Day, Rep, Par1, Par2) on the five responses (panels a-e).

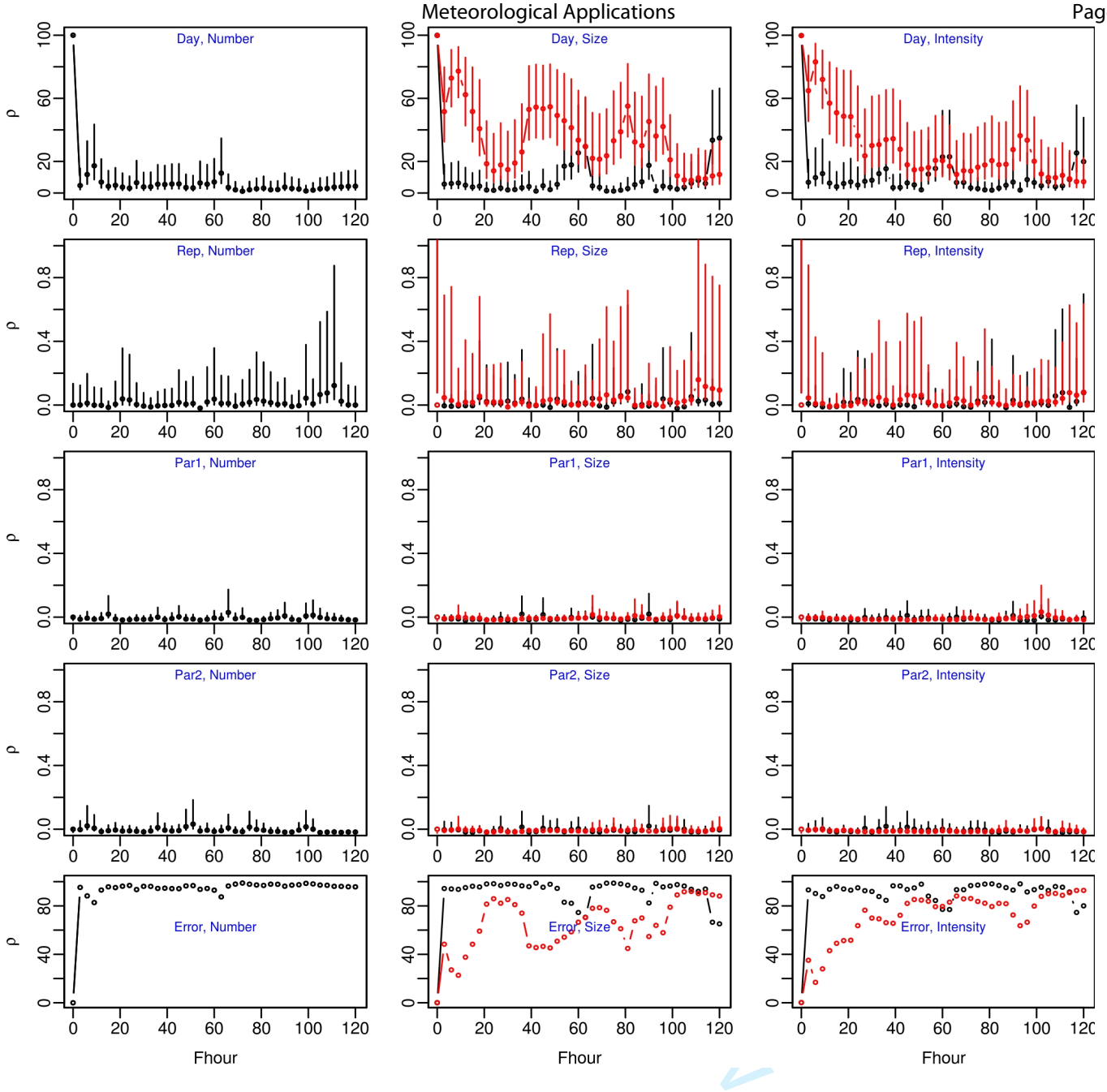

Figure 6. The 95% confidence intervals for the intraclass correlation ρ versus forecast time, displaying the effect of the four factors, Day, Rep, Par1 and Par2 (top 4 rows) on the five responses - number of objects (left column), minimum (black) and maximum (red) size (middle column), and minimum (black) and maximum (red) intensity (right column). The last row shows ρ_ϵ, the proportion of total variance in each response not explained by the four factors.

Table 1. An example of an LSD involving three factors A, B, C, each taking three values (denoted by the indices 1, 2, 3).

	A1	A2	A3
B1	C1	C2	C3
B2	C2	C3	C1
B3	C3	C1	C2

A Methodology for Sensitivity Analysis of Spatial Features in Forecasts: The Stochastic Kinetic Energy Backscatter Scheme

Caren Marzban*, Robert Tardif, Natalia Hryniw, Scott Sandgathe

All numerical models have parameters whose values are often set in an *ad hoc* fashion, and so, it is important to asses how these parameters affect the output of the model. The output of many models often contain "objects" examples of which are shown in the figure below. This paper proposes a methodology for assessing how the model parameters affect specific features of such objects.