Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2017-273-AC3, 2018


Geoscientific Model Development Discussions

# *Interactive comment on* "On the Effect of Model Parameters on Forecast Objects" *by* Caren Marzban et al.

**Caren Marzban et al.**

marzban@stat.washington.edu

Received and published: 4 January 2018

Dear Reviewer 3,

Thank you for your review. Below, please find your original comments (denoted with a ">") and our responses.

> General comments

> Summary ...

> This is an interesting paper, and describes methodology for an important problem in sensitivity analysis, namely conducting SA for a spatial field of model responses. Furthermore, the clustering both addresses the nature of precipitation data (i.e., non-



smooth or non-continuous data) while also reducing the dimension of the problem (i.e., considering fixed clusters rather than a fine spatial grid). The paper is well-written and nicely motivates the work, however, additional detail should be given in the section describing the statistical model, and I think reorganization of Section 3 would greatly improve the presentation (see the Technical corrections below). Furthermore, I am concerned with the analysis methods, particularly the significance testing, and am worried that the way in which the results are presented might be misleading (see the Scientific comments for more details).

We agree with all of your general comments. See more detailed responses below.

> Scientific comments

> As a statistician, I will primarily comment on the statistical model and significance testing, leaving discussion on the experimental design of the sensitivity analysis and variables selected for analysis (i.e., latitude, longitude, intensity, area, orientation, and eccentricity) to more informed parties. In my opinion, the clustering approaches considered (GMM and DBSCAN) seem reasonable, and it was nice to see that results are robust to the clustering method used.

Agreed.

> My first concern has to do with the description of the MMR model as well as the treatment of daily replicates within this model. The authors present a generic description of a multiple linear regression model in Equation (1), but it would be helpful to more clearly describe the generalization to the multivariate multiple linear regression model that was actually used. If I am following everything correctly, the statistical model you actually use is Eqn for MMR with 3 responses (where min = minimum, med = median, and max = maximum), or, written in vector form, Eqn for MRR in vector form (1) for t = 1, . . . , 99 samples taken from the 11-dimensional parameter space. Presumably, you use the usual MMR assumption that the error vectors delta_t are independent and identically distributed as Normal with mean vector 0 and non-diagonal covariance ma-

trix Sigma (i.e., the elements of delta_t are correlated). Is this a correct characterization of the model?

The Reviewer's description of our model is correct, and we will be happy to include the additional details in the paper.

> In practice, you actually estimate the 3 X 11 beta coefficients from Equation (1) for each of six features and each of 40 days, presenting boxplots of the beta coefficients aggregated over the 40 daily replicates for each of the 3 X 11 X 6 combinations of feature summaries/input variables/features. (See below for a concern related to the boxplots.) This seems like an unnecessary complication to the analysis. As evidenced by your decision to keep only every third day (reducing your data from 120 days to 40 days) in order to remove temporal correlation, it seems to me that these 40 days could represent an ensemble of realizations for each of the 99 parameter settings. Thus, instead of fitting 40 separate MMR models for each of the 6 features, your model could instead be Eqn for MMR in vector form but across all days (2) for d = 1, ... , 40 days (I assume that x_jtd = x_jt for j = 1, ... , 11, i.e., that the input parameter settings are the same for each day). In other words, instead of using the 40 daily replicates to estimate the distribution of each beta coefficient, you could build this variation into the statistical model and directly estimate the variability of the coefficients, then calculating P-values or confidence intervals as required. This seems to be a more refined way to handle the daily replicates, especially since it seems that you are not concerned with how the beta coefficients vary across the different days.

Here the Reviewer is concerned over how daily variability is handled. In the paper, we developed an MMR for each of the 40 days, while the proposed model in Eq (2) above, would "average" over the daily variability. While the latter model may make sense from the perspective of a Statistician aiming to build a most parsimonious model, the fact is that in most SA applications daily variability is something that users want to see. As such, averaging over it is not desirable for practitioners. There is a third alternative - introducing a factor, denoted Day, on the right side of the model. In other words, in



the language of experimental design, one can block the Day factor. We have actually performed that analysis as well. There are pros and cons to that work.

In general, on the one hand, blocking the Day factor is expected to make it easier to detect a statistically significant effect in the other 11 beta coefficients (i.e., it can increase power). On the other hand, because of the restriction on randomization (hence, treating Day as block), one cannot rely on the tests of significance for a block (i.e., Day) effect. Even if one were to believe the p-value associated with the Day factor, it would be only one number! And that brings us back to what we said earlier, namely that in most applications users desire to see the daily variability.

Now, that is all generalities and expectations; but what about the problem at hand? As we said, we have actually done the analysis of including the Day factor in the model as a block. Some of the results are reasonable conclusive. For example, when the response is simply the domain average of the forecast (i.e., not object-based at all), we found that blocking the Day factor has no effect on the estimates of the other 11 beta coefficients. But when dealing with objects the results do not suggest any simple conclusion! For that reason, we decided to exclude it from the paper. However, if the Reviewer believes this is too important to ignore, we will be happy to discuss it (perhaps in an appendix, in order to not disrupt the flow of the paper).

> Secondly, I am concerned by the significance testing procedure and the presentation of results. First of all, your two-stage procedure for controlling Type I error seems ad hoc, particularly your qualitative approach to assessing individual significance in the second stage.

We are surprised by the Reviewer's opinion on the 2-stage procedure. Outside of the multiple-hypothesis-testing circles, it is *the* approach to testing. One begins with a single omnibus test, and only if it's rejected one proceeds to performing multiple tests. There are numerous articles advocating the wisdom in this practice, and we will be happy to include them in the paper.

> The omnibus test in the first stage is a good idea (although it would be helpful to have more details given on exactly what you have done - instead of simply providing citations),

The omnibus test we performed is an F-test (again, a standard choice). Is this the kind of detail the Reviewer is proposing?

> but you need to be careful about the multiple testing even after reducing the number of tests to 6 X 40 = 240. I appreciate that you have at least considered a Bonferroni adjustment, but you should think carefully about this choice: Bonferroni controls a family-wise error rate, implying that the collective conclusion of all tests is invalid if at least one Type I error is made. I don't think this is actually what you want - it seems to me that you simply want to control the number of Type I errors. As an alternative, you might consider the very simple procedure for controlling the rate of false discoveries (i.e., FDR) given in the classic paper by Benjamini and Hochberg (1995). Their simple procedure is remarkably powerful and could more appropriately address the multiple testing issue.

It is not clear to us which error rate - FWER or FDR - is more appropriate to control for gridded fields, so we shall report the results of both. However, let us point out that the choice of the controlled error rate has very little bearing on the majority of the results in the paper, because in spite of the prevalence of p-values very little hypothesis testing is actually performed. There are only a few places in the paper where we report counts of significant effects. The remainder of the conclusions are based on the visual assessment of boxplots; in this connection, please see our response below.

> Regardless, after you have conducted the omnibus test, you proceed to present box plots of the coefficient estimates, aggregated across the daily replicates. I think that such an aggregation of the coefficient estimates provides you with a sampling distribution of the true coefficient estimate - please correct me if this is not the right way to think about this.



It may be safer to call it the "empirical" sampling distribution. Even then, some may object to calling it a sampling distribution because sampling across days is hardly a random sample from a population. But, yes, these boxplots are intended to summarize some proxy for the sampling distribution of the respective regression coefficients.

> In any case, the aggregated coefficient estimates are most certainly not a posterior distribution of the true coefficient, which is what you would get from a fully Bayesian analysis. In this case, it is misleading to represent a sampling distribution with a box-plot: if the boxplot is skewed to the right, this does not mean that the distribution of the true coefficient is skewed to the right.

Given that we have no a priori reasons for believing that there should be a skew, we have no reason to choose anything other than a symmetric a priori pdf. As such, the skew in the boxplots does translate to a skew in the posterior pdf.

> Instead, you should represent sampling distributions using a confidence interval, which could be plotted as a box (with no whiskers) or a solid bar.

A confidence interval has two "drawbacks:" 1) It does not convey the shape of the underlying distribution - a useful quantity, and 2) It depends on a significance/confidence level (see next comment, below).

> Additionally, simply checking to see if boxplots overlap with zero is not an appropriate way to assess statistical significance: what significance level is being considered?

The Reviewer is correct in that boxplots alone are not sufficient for performing hypothesis testing - one also requires some kind of threshold, e.g., significance level. However, as we have indicated above and in the paper, in spite of the prevalence of p-values in the paper, we actually do very little hypothesis testing (i.e., rejecting/not-rejecting). This is intentional. Although some problems can benefit from a simple significant/not-significant summary of results, in the case of our problem, we believe it is more informative to display the empirical sampling distributions. Although this certainly introduces a

subjective/qualitative ingredient into the analysis, we believe that it displays the results in a more holistic manner, and therefore, is a more useful trade-off. This philosophy is in line with the policy that many journals and practitioners are following in that summarizing complex results in terms of a binary reject/no-reject decision, or a p-value, or a confidence interval, leads to loss of information. We are hoping that the Reviewer will see the benefits of this trade-off, but if necessary we are willing to superimpose some sort of confidence interval on the boxplots (or the alternative shaded-point plots proposed above).

> My suggestion would be to fold the daily replicates into the MMR as suggested in Equation (2), and calculate P-values for each of the 11 X 3 X 6 coefficients. Then, I would use the Benjamini and Hochberg procedure to identify statistically significant coefficients at a particular level alpha. Instead of the boxplots in Figures 2 and 3, I would recommend using points or bars to indicate the magnitude of the coefficient estimate and shading or masking to indicate which estimates are statistically significant.

As we have indicated, we are amenable to discussing the various ways in which daily variability can be handled, and reporting counts of significant effects based on both FWER and FDR control. We can also see the benefit of replacing the boxplots with something that shows each of the members in the boxplots. Although this will take some experimentation on our part, we will do it because it's a good idea.

> Technical corrections

> On a more technical note, I found the organization of Section 3 to be very confusing. I would suggest moving Sections 3(d) and 3(e) to immediately follow Section 3(a). In this case you will have already described the clustering and the features of interest before discussing the statistical model and significance testing. I would also recommend moving lines 161-170 into Section 3(e).

We were aware that there is some "back-and-forthing" in that section, but we believed that structure was a reasonable trade-off. However, if the Reviewer found it "very con-



fusing," then we will be happy to re-organize as suggested.

Thank you.

---

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-273, 2017.