

Model Tuning with Canonical Correlation Analysis

CAREN MARZBAN

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

SCOTT SANDGATHE

Applied Physics Laboratory, University of Washington, Seattle, Washington

JAMES D. DOYLE

Naval Research Laboratory, Monterey, California

(Manuscript received 31 July 2013, in final form 4 January 2014)

ABSTRACT

Knowledge of the relationship between model parameters and forecast quantities is useful because it can aid in setting the values of the former for the purpose of having a desired effect on the latter. Here it is proposed that a well-established multivariate statistical method known as canonical correlation analysis can be formulated to gauge the strength of that relationship. The method is applied to several model parameters in the Coupled Ocean–Atmosphere Mesoscale Prediction System (COAMPS) for the purpose of “controlling” three forecast quantities: 1) convective precipitation, 2) stable precipitation, and 3) snow. It is shown that the model parameters employed here can be set to affect the sum, and the difference between convective and stable precipitation, while keeping snow mostly constant; a different combination of model parameters is shown to mostly affect the difference between stable precipitation and snow, with minimal effect on convective precipitation. In short, the proposed method cannot only capture the complex relationship between model parameters and forecast quantities, it can also be utilized to optimally control certain combinations of the latter.

1. Introduction

The relationship between model parameters and forecast quantities is often complex, but knowledge of that relationship has both theoretical and practical consequences. The former can shed light on the underlying physics, and the latter can help in setting the values of the model parameters to have some desirable effect on the forecasts. Numerous attempts at statistically modeling such relationships have been made (Gombos and Hansen 2008; Hacker et al. 2011; Torn and Hakim 2008).

In a recent study, Marzban et al. (2014) examined the relationship between 11 model parameters in the Coupled

Ocean–Atmosphere Mesoscale Prediction System (COAMPS)¹ and each of four forecast quantities: 24-h accumulated 1) convective, 2) stable, 3) total precipitation, and 4) snow. The approach employed a variance-based sensitivity analysis (Marzban 2013), ideally suited to the situation where a single forecast quantity is of interest.

One limitation of that work is that it does not incorporate relationships that are known to exist between forecast quantities. For example, as shown below, model parameter values that lead to increased convective precipitation are generally associated with decreased stable precipitation. In a fully multivariate treatment of multiple model parameters and multiple forecast quantities, it is important to account not only for associations

Corresponding author address: Caren Marzban, Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322.

E-mail: marzban@stat.washington.edu

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

between model parameters and forecast quantities, but also between forecast quantities and between model parameters. Said differently, the sensitivity analysis method employed in Marzban et al. (2014) takes into account associations between model parameters and forecast quantities, but not within model parameters, nor within forecast quantities. In this sense, that analysis is not a truly multivariate analysis.²

Multivariate methods are generally divided into two groups which (in the realm of machine learning) are referred to as supervised and unsupervised (Bishop 1996). The defining feature of the former class is the division of the variables under study into a set of predictors and a set of responses, and the goal is to infer a relationship between the two sets. An example of such a method is regression (Draper and Smith 1998). By contrast, such a distinction is not made in unsupervised methods. Instead, the goal is to find a combination of variables that account for most of the variability in the data. Principal components analysis (PCA) is a well-known example of an unsupervised method (Abdi and Williams 2010; Jolliffe 2002). The method employed here, known as canonical correlation analysis (CCA), borrows from both approaches in that the task involves two sets of variables, but the main goal is to find combinations of variables in one set and combinations of variables in the second set, which are most correlated with one another. In the current application the two sets of variables are the model parameters and the forecast quantities.

CCA (Anderson 2003; Glahn 1968; Mardia et al. 1979) is suited to understanding the relationship between model parameters and forecast quantities because 1) it treats the two sets of variables symmetrically, and 2) it takes into account relationships within each set. The symmetry is desirable because it provides a wholistic assessment of the relationship in the same sense in which Pearson's correlation coefficient provides a summary of the relationship between two quantities. Indeed, CCA can be described as a multivariate generalization of the correlation coefficient (analysis). The second feature is important because the relationship between two sets of variables can be confounded by the relationships between the variables within each set. By contrast, a regression approach treats the two sets of variables asymmetrically in that the map from the model parameters to the forecast quantities is not the same as the map from forecast quantities to model parameters. Also, in regression, the relationship between response variables is not

incorporated into the analysis. As such, for the problem at hand, CCA is more appropriate than a strictly regression approach. CCA is also more appropriate than any unsupervised approach (e.g., PCA), because it is the relationship between two sets of distinct variables that is of interest—a defining feature of CCA. It is worth mentioning that in spite of the differences between CCA and regression, many multivariate techniques (including CCA) can be formulated within a multivariate regression framework (i.e., with multiple predictors and multiple responses) wherein the two sets of data have been transformed in certain ways (Tippett et al. 2008).

CCA has been widely used in the atmospheric sciences, but mostly for the purpose of identifying predictors of the primary modes of oscillations in the climate, such as El Niño–Southern Oscillation (ENSO; Barnston and Ropelewski 1992; Nicholls 1987), Pacific decadal oscillation (PDO; Livezey and Smith 1999), or the North Pacific Oscillation (NPO; Anderson and Maloney 2006). CCA has been used for the prediction of monsoon rainfall (Singh et al. 2012), and a probabilistic extension of CCA has been developed and applied to Pacific sea surface temperatures by Wilks (2014).

The outline of this paper is as follows: the next section briefly describes the dataset under study. To set the stage for CCA, the method section begins with an application of regression to the data at hand; that analysis not only introduces the notation for the remainder of the paper, but its results also serve to reduce the number of model parameters under investigation. The results of CCA, presented next, suggest that certain combinations of the model parameters are highly correlated with certain combinations of the forecast quantities. It is found that further diagnosis of these combinations provides useful guidance in better setting the values of the model parameters. The paper ends with a summary and a discussion of the results and of future work.

2. Data

The dataset used in this study is that used in Marzban et al. (2014). The experimental design underlying the data is described by Bowman, Sacks and Chang (1993), Sacks et al. (1989), Santner et al. (2003), and Welch et al. (1992). Briefly, Latin hypercube sampling (Cioppa and Lucas 2007; Marzban 2013; Marzban et al. 2014) is used to generate 99 values of 11 model parameters. The choice of the model parameters is based on Holt et al. (2011). The result is data on model parameters, often called the empirical region. For each of the 99 points in the empirical region, the atmospheric portion of COAMPS (Hodur 1997; Doyle et al. 2011; Jiang and Doyle 2009), version 4.2.2, is used to generate 24-h forecasts of three quantities:

²In Marzban et al. (2014), as well as in the current study, the values of the model parameters are selected in a manner that precludes any association between them.

TABLE 1. The 11 parameters studied in this paper. Also shown are the default values and the range over which they are varied. Kain–Fritsch (KF) (Kain and Fritsch 1993), planetary boundary layer (PBL), and lifting condensation level (LCL).

Name	Description	Default	Range
mixlen	Linear factor that multiplies the mixing length within the PBL	1.0	0.5, 1.5
sfclx	Linear factor that modifies the surface fluxes	1.0	0.5, 1.5
wfctKF	Linear factor for the vertical velocity (grid scale) used by KF trigger	1.0	0.5, 1.5
delt1KF (°C)	Temperature increment at the LCL for KF trigger	0	-2, 2
delt2KF (°C)	Another method to perturb the temperature at the LCL in KF	0	-2, 2
prcfrac	Fraction of available precipitation in KF, fed back to the grid scale	0.5	0, 1
cloudrad (m)	Cloud radius factor in KF	1500	500, 3000
autocon1 (kg m ⁻³ s ⁻¹)	Autoconversion rate coefficient for the microphysics	0.001	1 × 10 ⁻⁴ , 1 × 10 ⁻²
autocon2 (kg m ⁻³ s ⁻¹)	Autoconversion mass threshold value	4 × 10 ⁻⁴	4 × 10 ⁻⁵ , 4 × 10 ⁻³
rainsi (m ⁻¹)	Slope intercept parameter for rain in the microphysics	8.0 × 10 ⁶	8.0 × 10 ⁵ , 8.0 × 10 ⁷
snows_i (m ⁻¹)	Slope intercept parameter for snow in the microphysics	2.0 × 10 ⁷	2.0 × 10 ⁶ , 2.0 × 10 ⁸

accumulated 1) convective precipitation, 2) stable (or grid scale) precipitation, and 3) snow.³ The forecasts are generated for each of 36 dates, beginning with 1 January and ending with 4 July 2009, at approximately 4-day intervals to assure independence. For each of the 99 model parameter values, and for each date, the 90th percentile (across the spatial domain) of the forecast quantities is computed. These quantities (measured in mm) constitute the forecast quantities of interest in this work. In other words, the focus of the current study is on “heavy” precipitation and snow. Here, these three forecast quantities are denoted by the symbols conv, stab, and snow, respectively. The model parameters are shown in Table 1.

The COAMPS model was forced using 0.5° resolution initial and one-way boundary conditions from the Navy Operational Global Atmospheric Prediction System (NOGAPS). The COAMPS analysis domain is a 72 × 45 grid, roughly covering the continental United States, and the fields are based on NOGAPS initial fields and local observations. Given that the focus of the CCA method is inferring a map relating forecast quantities and model parameters—not forecast quality—for computational efficiency COAMPS is run at a resolution of 81 km.

3. Method

As mentioned previously, one of the reasons for considering CCA for inferring the relationship between model parameters and forecast quantities is that it also incorporates relationships between forecast quantities. That the relationship exists at all can be seen in the scatterplot between the forecast quantities. Figure 1 shows the scatterplot of stable precipitation versus convective

precipitation for four dates sampled from across the period available in the dataset. Each panel contains 99 points corresponding to the 99 points in the empirical region (i.e., each point corresponds to a different value assigned to the 11 model parameters). These four dates are selected to illustrate some common features. For 1 January, the relationship is weakly linear, with a negative slope. A stronger negative association can be seen on 6 February; and on some days (e.g., 18 March) there is no association between the two forecast quantities at all. Some of these patterns repeat on different days. For example, the pattern seen on 13 May is nearly identical to that of 1 January. Although not shown here, on majority of the days the relationship is of the type found on 1 January and 6 February (i.e., linearly and negatively associated). In the span of dates examined here, on no day is a positive association observed. In other words, model parameter settings that lead to increased convective precipitation are generally associated with decreased stable precipitation. Similar patterns exist between convective precipitation and accumulated snow, although the associations are much weaker than those shown in Fig. 1. It is such linear relationships that CCA takes into account in identifying combinations of forecast quantities which are most correlated with combinations of model parameters.

Before introducing the details of CCA, it is useful to consider multivariate regression. Let the three forecast quantities be denoted by y_j , where $j = 1, 2, 3$, and the model parameters by x_i , where $i = 1, \dots, 11$. Furthermore, assume that the data on x_i and y_j have been standardized to have zero mean and a standard deviation of 1.⁴ First, the following multivariate linear regression model is used

³Total precipitation, analyzed in Marzban et al. (2014), is not analyzed directly here because it is simply the sum of convective and stable precipitation.

⁴For any quantity u , such a standardization is obtained by $(u_k - \bar{u})/s$, where u_k is the k th observation of u ; and \bar{u} and s are the sample mean and sample standard deviation of u , respectively.

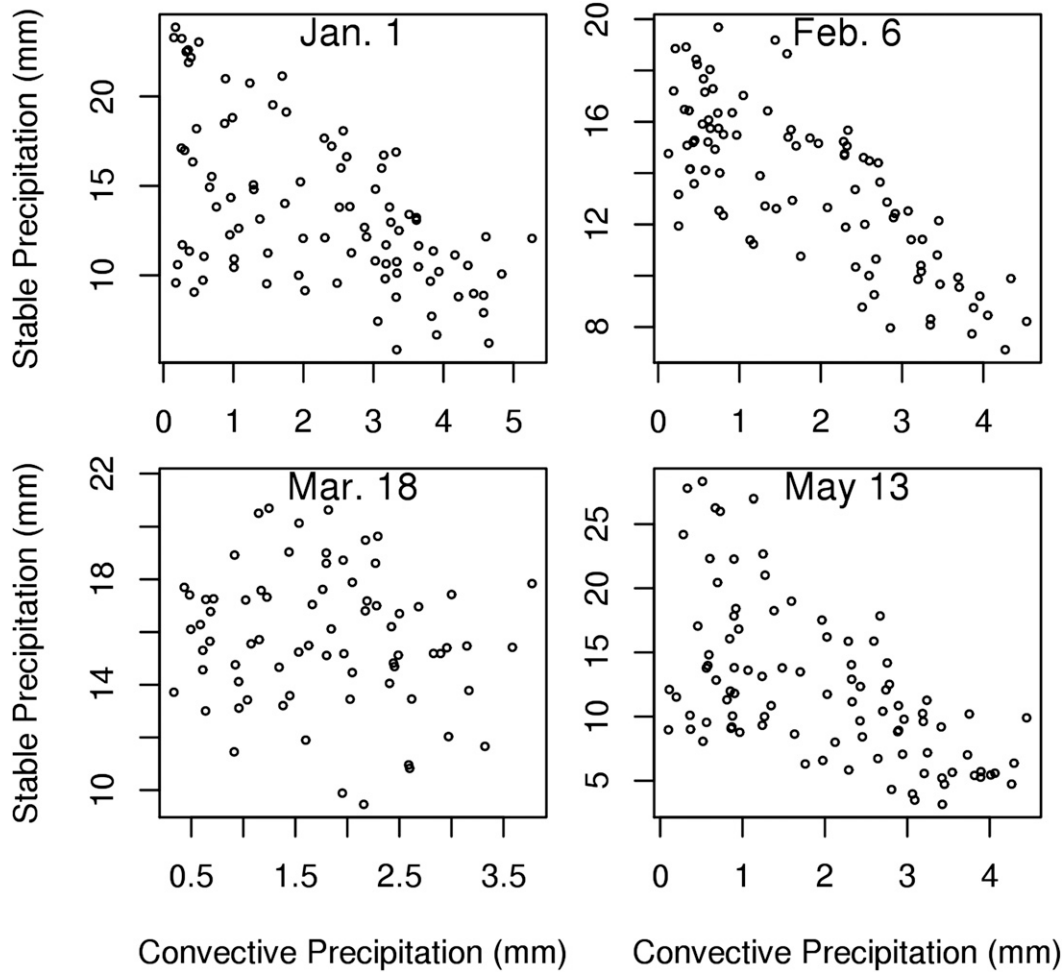


FIG. 1. Scatterplot of stable precipitation vs convective precipitation, across 99 points in parameter space, for four different days.

to represent the relationship between the x_i and the y_j (Hastie et al. 2001):

$$y_j = \beta_{0j} + \sum_{i=1}^{11} \beta_{ji} x_i + \epsilon_j, \quad j = 1, 2, 3, \quad (1)$$

where the errors ϵ_j are assumed to have normal distributions with zero mean; for clarity, in this equation the case index is not shown. Estimates of the β coefficients provide a measure of the importance of the corresponding x_i in terms of its effect on the y_j . Interpreting regression coefficients in this manner can be problematic if the predictors are collinear; but this is not a problem in the current study because the model parameters are sampled (Marzban et al. 2014) in a way to assure that there is no collinearity.

It can be shown (Hastie et al. 2001, p. 54; Tippett et al. 2008) that the minimization of mean-squared error for the model in Eq. (1) actually leads to estimates of the β

parameters as if they were estimated via three separate regression fits, one for each of y_j . As such, the model is inadequate in the sense that it does not incorporate the relationships between forecast quantities. However, such a model is still useful in that it allows one to identify which parameters have no affect at all on any of the forecast parameters. As shown below, this model is used to reduce the number of model parameters from 11 to 8.

A fully multivariate analysis is provided by CCA. In fact, the basic quantities of CCA are linear combinations of x_i and linear combinations of y_j . They are referred to as canonical variates (CV) of x and y , respectively—or CV pairs, in general. The goal of CCA (in its simplest form) is to find CV pairs with the highest possible Pearson's correlation coefficient. When such a pair exists, the coefficients in the respective CVs—called loadings—measure the contribution of the corresponding variable to that CV in the same way in which the β coefficients in the regression model in Eq. (1) measure the contribution

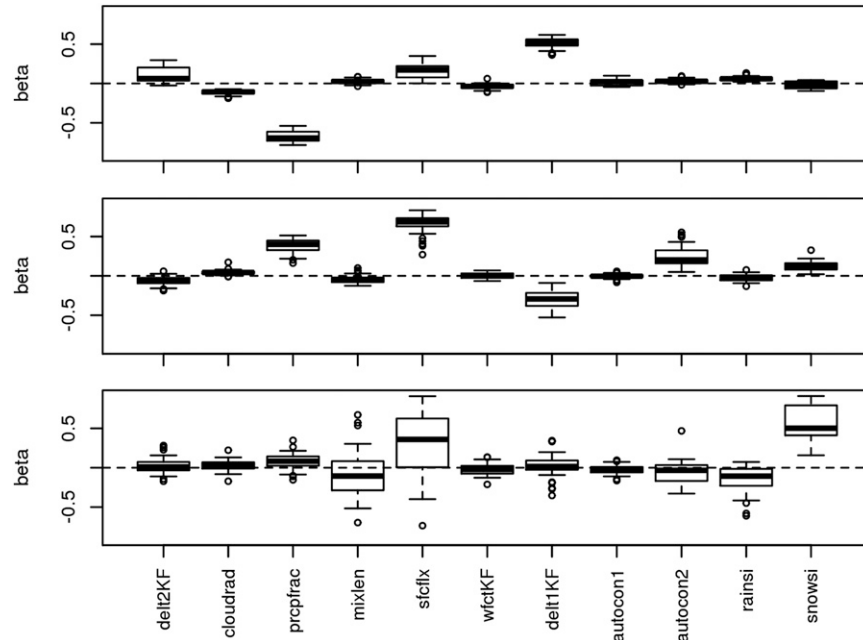


FIG. 2. The distribution, across 36 days, of the regression coefficients [β in Eq. (1)] associated with 11 model parameters (along x axis) when the response is (top) convective precipitation, (middle) stable precipitation, and (bottom) snow.

of x_i to y_j . For example, a loading of 0.6 for a given model parameter means that a change of one standard deviation in that model parameter is expected to lead to an average change of 0.6 in the corresponding CV. Similarly, a loading of 0.7 for a given forecast quantity means that a change of one standard deviation in that quantity is expected to lead to an average change of 0.7 in its CV. Examination of the loadings on the CV pairs often allows one to associate a physical meaning to the quantity represented by the CV pairs. As in PCA, the interpretation of the loadings is not unique (Jolliffe 2002), but that is not a problem here because our focus is not on identifying a physical underlying process. The main aim here is to identify the CV pairs for the purpose of optimally controlling the forecast quantities; a unique interpretation of the loadings is not necessary for that purpose.

CCA is performed on eight model parameters (selected based on the aforementioned multivariate regression analysis) and the three forecast quantities, for each of the 36 days in the dataset. The distribution, across days, of the loadings is then summarized by boxplots. If the boxplot of the loading for a given model parameter is centered near zero, then that model parameter can be considered unimportant in terms of its reliability in affecting the forecast parameters, because on the average (across days) the loading is near zero. Here, no objective criterion is employed to decide whether or not a boxplot

is “centered near zero,” and as such the interpretation of these boxplots is necessarily qualitative; but this approach does have the advantage of displaying the daily variability of the effect of the model parameters on forecast quantities.

For instance, a relatively wide boxplot for the loading of a given model parameter implies large daily variability, and so, that model parameter can be considered unreliable because its effect on the forecast quantities will be inconsistent across days. Although, an attempt is made here to maintain the qualitative nature of the conclusions, at times an appeal to some criterion is made in order to simplify the conclusions; for example, if zero falls within the interquartile range (i.e., within the box of the boxplot), then the corresponding variable is considered to be unimportant.

4. Results

Figure 2 shows the distribution of the β regression coefficients in Eq. (1), for each of the forecast quantities. Each boxplot summarizes the distribution (across 36 days) of the regression coefficients corresponding to the model parameters.⁵ It can be seen that convective precipitation

⁵The circles denote outliers, conventionally defined by any case beyond 1.5 times the interquartile range of the mean.

(top panel) is affected mostly by the fraction of available precipitation in Kain–Fritsch (KF) fed back to the grid scale (prcpfrac) and the temperature increment at the lifted condensation level (LCL) for the KF trigger (delt1KF). Specifically, and as expected, an increase in the former (latter) is accompanied by a decrease (increase) in convective precipitation. The model parameter associated with a different method to perturb the temperature at the LCL in KF (delt2KF), the cloud radius factor in KF (cloudrad), and a linear factor that modifies the surface fluxes (sfcflx) also affect convective precipitation, but to a much lower degree. The remaining model parameters have little or no effect on convective precipitation. These conclusions are consistent with those found in Marzban et al. (2014).

As seen in Fig. 2 (middle panel), stable precipitation is most affected by sfcflx, followed by prcpfrac, delt1KF, and the autoconversion factor for the microphysics (autocon2), and to a far lower degree by the slope intercept parameter for snow in the microphysics (snowsi).

As for snow (Fig. 2, bottom panel), the relatively large daily variability in the regression coefficients, reflected in the size/spread of the boxplots, suggests that snow is a more difficult forecast quantity to control with model parameters. It appears that the snow slope intercept parameter is the only of the 11 parameters with an unambiguous effect on snow. A linear factor that multiplies the mixing length within the PBL (mixlen), the surface flux factor (sfcflx), and the slope intercept parameter for rain in the microphysics (rainsi) all have marginal effects on snow either because of the large variability of the regression coefficients, or because zero falls within the interquartile range of the distribution.

Although the 11 model parameters have a complex relationship with the 3 forecast quantities, it appears that the parameters mixlen, wfctKF (a linear factor for the vertical velocity used by KF trigger), and autocon1 have little or no effect on any of the forecast quantities.⁶ Therefore, these parameters are henceforth excluded from analysis, reducing the number of model parameters from 11 to 8.

Note that an increase in the fraction of available precipitation (prcpfrac) is associated with a decrease in convective precipitation (top panel) but an increase in stable precipitation (middle panel). Similarly, delt1KF is positively associated with convective precipitation but negatively associated with stable precipitation. In other

words, the effect of the parameters on the forecast quantities is complex. The CVs constructed in CCA are designed to incorporate such relationships.

In applying CCA to the data at hand, the first question is (in admittedly poor English) “what linear combination of model parameters is most correlated with what linear combination of forecast quantities?” Given that there are only three forecast quantities present, only three such linear combinations can be formed. The linear combinations (i.e., the CVs) and the coefficients in each linear combination (i.e., the loadings) are the central entities in CCA.

By design, the largest correlation coefficient is between the first CV of the model parameters and the first CV of the forecast quantities. For the first day in the data set it is 0.936. The analogous correlations for the second and third CV pairs are 0.918, and 0.832. A Wilk’s Lambda test (Knapp 1978) of all three correlations leads to near-zero p values, implying that these correlations are statistically significant. The histogram (across 36 days) of the three correlations is shown Fig. 3. It is clear that all of the correlations are relatively large; the corresponding p values (not shown) are all near-zero. This suggests that there exist linear combinations of the model parameters which are highly correlated with linear combinations of the forecast quantities. A scatterplot of the CV of y versus the CV of x , across the 99 cases, displays a linear relationship for each of the three CVs, and for each of the 36 days. As such, it appears to be sufficient to examine only linear combinations of the model parameters and of the forecast quantities.

The loadings for the first CV of the model parameters and the first CV of the forecast quantities are shown in the top row of Fig. 4. Evidently, the first CV of the model parameters is mostly a measure of the surface flux factor (sfcflx). With the exception of the temperature perturbation parameter (delt2KF) and the slope intercept parameter for rain (rainsi), which have no contribution to the first CV, the fraction of available precipitation parameter (prcpfrac), the autoconversion mass threshold (autocon2), and the slope intercept parameter for snow (snowsi), all appear to have some contribution to the first CV, but to a much lower degree than sfcflx. The cloud radius factor (cloudrad) plays a unique role in that its boxplot is mostly below the horizontal line at zero, but only nearly so. In other words, that parameter appears to have a nonzero, albeit small, contribution to the first CV of the model parameters. As seen in the top-right panel in Fig. 4, the first CV of the forecast quantities essentially measures the sum of convective and stable precipitation. The large daily variability of the loading for snow implies that its contribution to the first CV is highly variable across days, and in that sense unreliable. To first

⁶The manner in which these parameters are found to be unimportant assumes that they do not interact with other parameters. Marzban et al. (2013) addressed the issue of interactions and found that they are statistically nonsignificant.

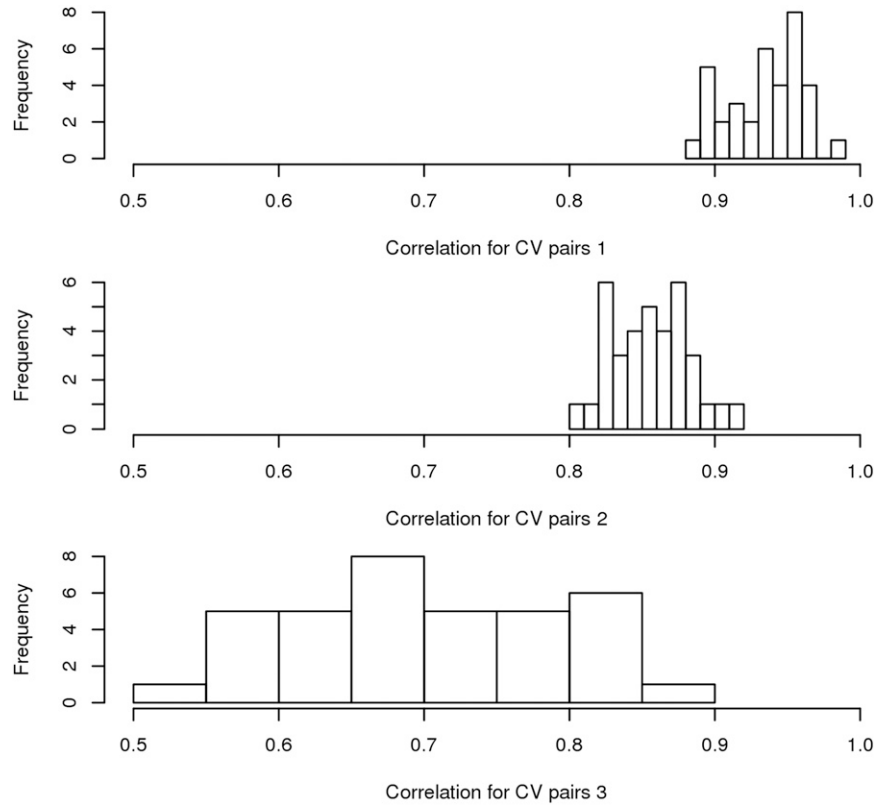


FIG. 3. The histogram, across 36 days, of the correlation coefficients between CV pairs.

approximation, therefore, the surface flux factor (sfcflx) alone is the best parameter for controlling total precipitation. Specifically, an increase of about one standard deviation in sfcflx is expected to increase total precipitation by approximately 0.8 standard deviations. An even more efficient way of increasing total precipitation would involve additionally decreasing the fraction of available precipitation (prcpfrac), and increasing autocon2 and snowsi, all by about 0.2 standard deviations. All of these standard deviation values are approximate and refer to the median of the boxplots.

The second CV of the model parameters mostly represents the difference between prcpfrac and delt1KF (Fig. 4, middle-left panel), because these parameters have the largest loadings, and appear with opposite signs. By similar reasoning, the second CV of the forecast quantities is mostly a representation of the difference between convective and stable precipitation (Fig. 4, middle-right panel). This is useful, because it implies that if one desires to increase convective precipitation and simultaneously decrease stable precipitation, while not affecting snow appreciably, then the best way is to decrease prcpfrac and increase delt1KF simultaneously. A more effective way of having the same effect on the forecast quantities would

involve also increasing delt2KF while decreasing sfcflx and snowsi. The magnitude of the changes can be determined from the y values in Fig. 4 (middle panels).

As shown in Fig. 4 (bottom-left panel), the third CV of the model parameters appears to be mostly affected by the snow slope intercept parameter (snowsi), and to a lesser degree by the autoconversion mass threshold parameter (autocon2) and the rain slope intercept parameter (rainsi). At first approximation, the third CV of the forecast quantities is dominated by snow (Fig. 4, bottom-right panel). As such, the best way of increasing snow is to increase snowsi. Examining the next level of contributions to the CV pairs, autocon2 and rainsi do have some contribution to the CV of the parameter values; and the CV for the forecast quantities appears to have a relatively large (and negative) loading on the two types of precipitation. All of these boxplots show large daily variability, and therefore, the effect of the model parameters on the forecast quantities is likely to be highly variable across days. On the average, however, a decrease in autocon2 and rainsi (in addition to an increase in snowsi), is expected to lead to an increase in snow and a simultaneous (but variable) decrease in total precipitation.

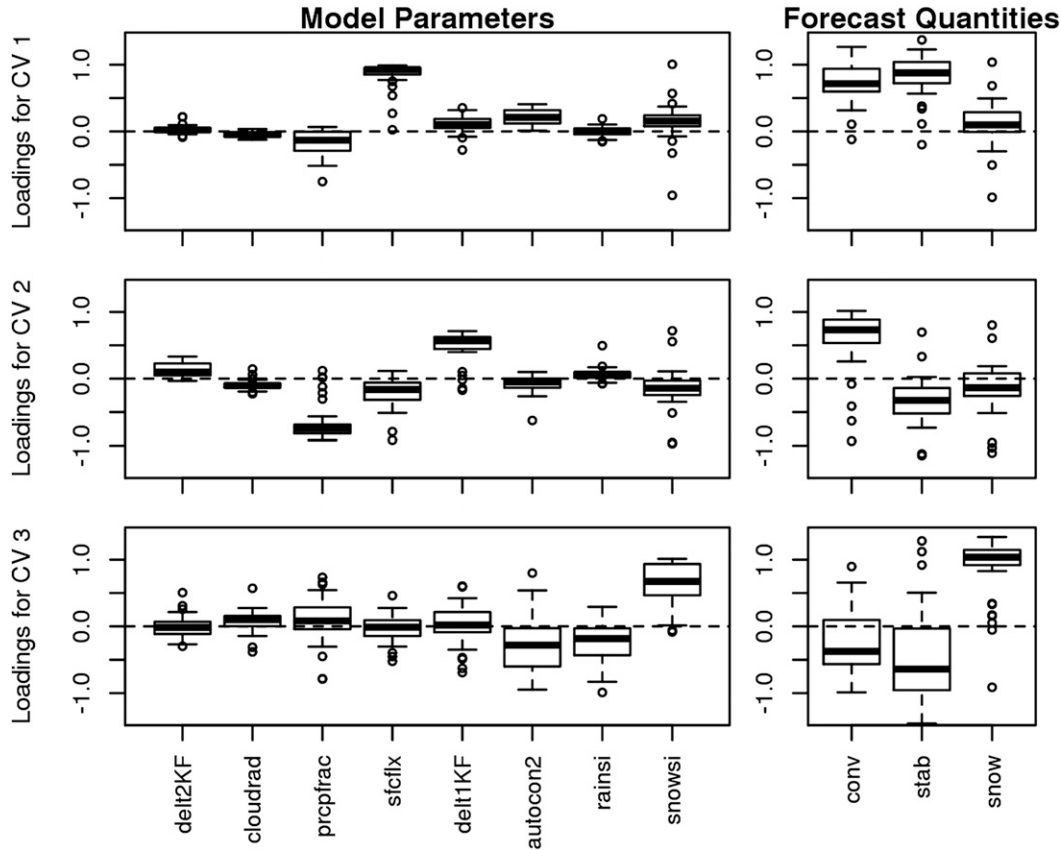


FIG. 4. The distribution, across 36 days, of the loadings (left) on the model parameters and (right) on the forecast quantities for the three CV pairs. The circles denote outliers (defined in text).

5. Summary and discussion

The effect of model parameters on forecast quantities is examined through multivariate statistical techniques. First, multivariate regression (i.e., with multiple predictors and multiple responses) is used to show that the linear factor that multiplies the mixing length within the PBL (mixlen), the vertical velocity factor (wfctKF), and the autoconversion factor for the microphysics (autocon1) have little to no effect on any of the forecast parameters examined—convective precipitation, stable precipitation, and snow. Eight other model parameters are found to have some type of effect on the forecast quantities. The relationship is found to be complex in that no single model parameter individually controls a single forecast quantity. Then CCA is employed to test whether any combination of forecast quantities and any combination of model parameters are well correlated. It is found that such combinations (i.e., CVs) do exist. The most-correlated CV pairs have correlation coefficients in the 0.88–0.98 range across 36 days. Two other CV pairs have correlation coefficients in the 0.81–0.92, and

0.61–0.86 range, respectively. All of these correlations are statistically significant with near-zero p values.⁷

A qualitative analysis of the contribution (i.e., loadings) of the model parameters and the forecast quantities to their respective CVs suggests several conclusions, with varying levels of complexity. At the simplest level, the surface flux factor (sfcflx) alone is responsible for controlling total precipitation (convective plus stable), while leaving snow mostly unaffected. By contrast, an increase in the difference between convective and stable precipitation is best obtained by a decrease in the fraction of available precipitation (prcpfrac) and an increase in the temperature increment at the LCL for KF trigger (delt1KF). Finally, an increase in snow alone is best accomplished through an increase in the slope intercept parameter for snow in the microphysics (snowsi). More complex relationships are also present, but at a weaker level.

⁷The largest p value for any CV pair and on any day is 0.00045.

It is important to recall that some parameters are found to be unimportant because of the large daily variability associated with them. It is possible that these parameters are in fact important (i.e., have a significant effect on forecasts), but only for certain meteorological events (e.g., fronts). It will be worthwhile to repeat the analysis performed here, but on datasets partitioned according to weather types. A geographic partitioning of the data may also be useful in revealing any spatial dependence of the results.

The above approach identifies certain combinations of the forecast quantities that are most affected by the model parameters. A more useful approach would allow one to control *any* combination of the forecast quantities. For example, one may desire to increase stable precipitation, while keeping convective precipitation and snow constant. That particular combination does not arise in the CCA performed here. Although it is possible that the inclusion of other model parameters can allow one to control other combinations of the forecast quantities, a more direct method would be to simply model the relationship between x_i and y_j , but using the y_j (forecast quantities) as predictors and x_i (model parameters) as responses. In such a model, one can find the optimal combination of the model parameters for any desirable outcome on the forecast quantities. A multivariate regression model of the type in Eq. (1), with x_i and y_j switched, will not be adequate because, as mentioned before, such a multivariate regression model is equivalent to a system of independent single-response regression models. Such a model will, therefore, not incorporate the relationship between the forecast quantities. CCA will also be inadequate, because it is inherently symmetric with respect to the x_i and y_j variables. Examining linear combinations of the CVs themselves will also not allow one to control each forecast quantity separately. Alternative approaches that will allow one to implement any desirable effect on the forecast quantities are currently under investigation.

In the current application of CCA, it has been employed to identify combinations of predictors and combinations of responses that are most correlated. In this sense, CCA is similar to two PCAs: one on the predictors and another on the responses. Also, as in PCA, one can generate predictions not only for the CVs, but also for the physical response variables. The predictive aspect of CCA has been emphasized by Glahn (1968) and Wilks (2014). Here, the predictive facet of CCA has not been utilized, because focus has been placed on the task of identifying the aforementioned combinations, and their daily variability. The quality of the resulting models has been assessed in terms of the correlation coefficient between the CVs. Viewing a CCA model as a prediction

model allows for assessing the quality of the model by comparing its predictions with observations. Although the results of that analysis are not presented here, it is found that the correlation coefficient between observations and predictions is generally in the 0.5–0.9 range, with the exception of a few days for which the correlation coefficients are very low (<0.4). A more complete assessment using verification methods may reveal why CCA is not a good model for some of the days in the dataset examined here.

Another possible generalization of this work is to allow for nonlinear relationships between the model parameters and forecast quantities, and/or allow interactions within each set. There exist multiple ways of incorporating nonlinear relations, ranging from the more traditional (Luijters et al. 1994) to more recent approaches based on neural networks (Cannon and Hsieh 2008; Hsieh 2000). Modeling nonlinear relationships is generally more complex than that of linear relationships, because care must be taken to avoid overfitting. More data are also required, and so, that work will be considered in the future.

Acknowledgments. Partial support for this project was provided by the Office of Naval Research Grants N00014-01-G-0460/0049 and N00014-05-1-0843. C.M. thanks useful conversations with William Hsieh, and J.D.D. acknowledges the support of the Chief of Naval Research through Program Element 0601153N of the Naval Research Laboratory Base Program.

REFERENCES

- Abdi, H., and L. J. Williams, 2010: Principal component analysis. *Wiley Interdiscip. Rev.: Comput. Stat.*, **2**, 433–459, doi:10.1002/wics.101.
- Anderson, B. T., and E. Maloney, 2006: Interannual tropical Pacific sea surface temperatures and their relation to preceding sea level pressures in the NCAR CCSM2. *J. Climate*, **19**, 998–101, doi:10.1175/JCLI3674.1.
- Anderson, T. W., 2003: *An Introduction to Multivariate Statistical Analysis*. 3rd ed. Wiley-Interscience, 752 pp.
- Barnston, A. G., and C. F. Ropelewski, 1992: Prediction of ENSO episodes using canonical correlation analysis. *J. Climate*, **5**, 1316–1345, doi:10.1175/1520-0442(1992)005<1316:POEEUC>2.0.CO;2.
- Bishop, C. M., 1996: *Neural Networks for Pattern Recognition*. Clarendon Press, 482 pp.
- Bowman, K. P., J. Sacks, and Y.-F. Chang, 1993: Design and analysis of numerical experiments. *J. Atmos. Sci.*, **50**, 1267–1278, doi:10.1175/1520-0469(1993)050<1267:DAAONE>2.0.CO;2.
- Cannon, A., and W. W. Hsieh, 2008: Robust nonlinear canonical correlation analysis: Application to seasonal climate forecasting. *Nonlinear Processes Geophys.*, **15**, 221–232, doi:10.5194/npg-15-221-2008.
- Cioppa, T., and T. Lucas, 2007: Efficient nearly orthogonal and space-filling Latin hypercubes. *Technometrics*, **49**, 45–55, doi:10.1198/004017006000000453.
- Doyle, J. D., Q. Jiang, R. B. Smith, and V. Grubii, 2011: Three-dimensional characteristics of stratospheric mountain waves

- during T-REX. *Mon. Wea. Rev.*, **139**, 3–23, doi:10.1175/2010MWR3466.1.
- Draper, N. R., and H. Smith, 1998: *Applied Regression Analysis*. 3rd ed. Wiley-Interscience, 736 pp.
- Glahn, H. R., 1968: Canonical correlation and its relationship to discriminant analysis and multiple regression. *J. Atmos. Sci.*, **25**, 23–31, doi:10.1175/1520-0469(1968)025<0023:CCAIRT>2.0.CO;2.
- Gombos, D., and J. A. Hansen, 2008: Potential vorticity regression and its relationship to dynamical piecewise inversion. *Mon. Wea. Rev.*, **136**, 2668–2682, doi:10.1175/2007MWR2165.1.
- Hacker, J. P., C. Snyder, S.-Y. Ha, and M. Pocerlich, 2011: Linear and non-linear response to parameter variations in a mesoscale model. *Tellus*, **63A**, 429–444, doi:10.1111/j.1600-0870.2010.00505.x.
- Hastie, T., R. Tibshirani, and J. Friedman, 2001: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 533 pp.
- Hodur, R. M., 1997: The Naval Research Laboratory's Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS). *Mon. Wea. Rev.*, **125**, 1414–1430, doi:10.1175/1520-0493(1997)125<1414:TNRLSC>2.0.CO;2.
- Holt, T. R., J. A. Cummings, C. H. Bishop, J. D. Doyle, X. Hong, S. Chen, and Y. Jin, 2011: Development and testing of a coupled ocean-atmosphere mesoscale ensemble prediction system. *Ocean Dyn.*, **61**, 1937–1954, doi:10.1007/s10236-011-0449-9.
- Hsieh, W. W., 2000: Nonlinear canonical correlation analysis by neural networks. *Neural Networks*, **13**, 1095–1105, doi:10.1016/S0893-6080(00)00067-8.
- Jiang, Q., and J. D. Doyle, 2009: The impact of moisture on mountain waves. *Mon. Wea. Rev.*, **137**, 3888–3906, doi:10.1175/2009MWR2985.1.
- Jolliffe, I. T., 2002: *Principal Component Analysis*. 2nd ed. Springer, 489 pp.
- Kain, J. S., and J. M. Fritsch, 1993: Convective parameterization for mesoscale models: The Kain-Fritsch scheme. *The Representation of Cumulus Convection in Numerical Models*, Meteor. Monogr., No. 46, Amer. Meteor. Soc., 165–170.
- Knapp, T. R., 1978: Canonical correlation analysis: A general parametric significance-testing system. *Psychol. Bull.*, **85**, 410–416, doi:10.1037/0033-2909.85.2.410.
- Livezey, R. E., and T. M. Smith, 1999: Covariability of aspects of North American climate with global sea surface temperatures on interannual to interdecadal timescales. *J. Climate*, **12**, 289–302, doi:10.1175/1520-0442-12.1.289.
- Luijckens, K., F. Symons, and M. Vuylsteke-Wauters, 1994: Linear and non-linear canonical correlation analysis: An exploratory tool for the analysis of group-structured data. *J. Appl. Stat.*, **21**, 43–61, doi:10.1080/757583648.
- Mardia, K. V., J. T. Kent, and J. M. Bibby, 1979: *Multivariate Analysis*. Academic Press, 521 pp.
- Marzban, C., 2013: Variance-based sensitivity analysis: An illustration on the Lorenz'63 model. *Mon. Wea. Rev.*, **141**, 4069–4079, doi:10.1175/MWR-D-13-00032.1.
- , S. Sandgathe, J. D. Doyle, and N. C. Lederer, 2014: Variance-based sensitivity analysis: Preliminary results in COAMPS. *Mon. Wea. Rev.*, in press.
- Nicholls, N., 1987: The use of canonical correlation to study teleconnections. *Mon. Wea. Rev.*, **115**, 393–399, doi:10.1175/1520-0493(1987)115<0393:TUOCCT>2.0.CO;2.
- Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn, 1989: Design and analysis of computer experiments. *Stat. Sci.*, **4**, 409–423, doi:10.1214/ss/1177012413.
- Santner, T. J., B. J. Williams, and W. I. Notz, 2003: *The Design and Analysis of Computer Experiments*. Springer, 299 pp.
- Singh, A., M. A. Kulkarni, U. C. Mohanty, C. Kar, A. W. Robertson, and G. Mishra, 2012: Prediction of Indian summer monsoon rainfall (ISMR) using canonical correlation analysis of global circulation model product. *Meteor. Appl.*, **19**, 179–188, doi:10.1002/met.1333.
- Tippett, M. K., T. DelSole, S. J. Mason, and A. G. Barnston, 2008: Regression-based methods for finding coupled patterns. *J. Climate*, **21**, 4384–4398, doi:10.1175/2008JCLI2150.1.
- Torn, R. D., and G. Hakim, 2008: Ensemble-based sensitivity analysis. *Mon. Wea. Rev.*, **136**, 663–677, doi:10.1175/2007MWR2132.1.
- Welch, W. J., R. J. Buck, J. Sacks, H. P. Wynn, T. J. Mitchell, and M. D. Morris, 1992: Screening, predicting, and computer experiments. *Technometrics*, **34**, 15–25, doi:10.2307/1269548.
- Wilks, D. S., 2014: Probabilistic canonical correlation analysis forecasts, with application to tropical Pacific sea-surface temperatures. *Int. J. Climatol.*, doi:10.1002/joc.3771, in press.

AU2