AUTOMATIC MODEL CALIBRATION

A New Way to Improve Numerical Weather Forecasting

Q. Duan, Z. Di, J. Quan, C. Wang, W. Gong, Y. Gan,
A. Ye, C. Miao, S. Miao, X. Liang, and S. Fan

Automatic model calibration improves numerical weather forecasting by tuning the numerical weather prediction model parameters to match model predictions with observations.

Weather forecasting skill has been improving steadily over the years. The improvement is due mostly to advances in the representation of physical processes by numerical weather prediction (NWP) models, observational systems, analyses techniques and forecasting methods (e.g., data assimilation, statistical postprocessing, and ensemble forecasting), new computational capability, and effective communications and training. There is an area that has received less attention so far but can bring significant improvement to weather forecasting—the calibration of NWP models. Model calibration refers to a process in which model parameters are tuned to match model predictions with corresponding observations. This process may be done using a manual, unsystematic "trial and error" approach, or using an optimization algorithm to tune model parameters automatically to minimize the difference between model predictions and observations (Duan et al. 2006). Automatic model calibration is a common practice in many fields including hydrology, biology, communications, and finance. It focuses on reducing errors resulting from the specification of model parameters. This is different from other popular approaches. For example, a common approach is to reduce model structural error by developing better physical parameterization schemes (Stensrud 2007). Multimodel ensemble approaches have been employed to account for model structural uncertainty (Krishnamurti et al. 1999). Data assimilation methods are commonly used to reduce errors in model initial conditions (Kalnay 2003).

There are several reasons that the automatic calibration of NWP models is not practiced as widely as in other fields. First, a typical NWP model has many

AFFILIATIONS: Duan, Di, Quan, Wang, Gong, Ye, and C. Miao—State Key Laboratory of Earth Surface Processes and Resource Ecology, Faculty of Geographical Science, Beijing Normal University, Beijing, China; Gan and Liang—State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, Beijing, China; S. Miao, and Fan—Institute of Urban Meteorology, China Meteorological Administration, Beijing, China
CORRESPONDING AUTHOR: Qingyun Duan, qyduan@bnu.edu.cn

The abstract for this article can be found in this issue, following the table of contents.
DOI:10.1175/BAMS-D-15-00104.1

A supplement to this article is available online (10.1175/BAMS-D-15-00104.2)

In final form 24 August 2016

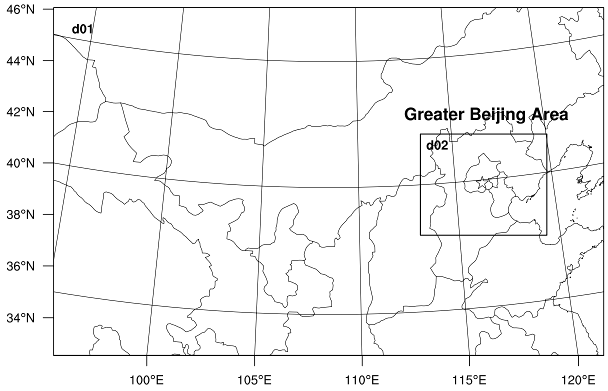

Fig. 1. The two-level nested study domain for the WRF Model. The outer domain is a 60 × 48 grid with a resolution of 27 km, and the inner domain is an 87 × 55 grid with a resolution of 9 km.

parameters (from tens to hundreds), which appear in model equations as constants or exponents. Parameter values may vary according to local conditions or climate regimes and are usually not measurable at the scale of application. The large number of parameters leads to a "curse of dimensionality" problem and makes optimization processes intractable. Second, NWP models simulate many meteorological variables, including precipitation, air temperature, atmospheric pressure, humidity, and wind speed. Model calibration must ensure that the simulation of all key meteorological variables is satisfactory. This requires model calibration be done via a multiobjective approach, which further increases the complexity of model calibration. Third, conventional automatic model calibration approaches require many model runs (up to tens of thousands) to obtain optimal parameter values. Since many CPUs are required to generate a multiday forecast for a limited domain, the extraordinary computational demand makes automatic model calibration very challenging.

The utility of automatic calibration of NWP models has been suggested by a number of researchers (Bennett et al. 1996; Evensen et al. 1998; Duane and Hacker 2007; Aksoy 2015). Various approaches have been attempted to optimize NWP model parameters, from traditional search algorithms such as downhill simplex method (Severijns and Hazeleger 2005), genetic algorithm (Yu et al. 2013; Ihshaish et al. 2012), and simulated annealing (Jackson et al. 2004), to adaptive sequential data assimilation (Gong et al. 1998; Mu et al. 2002; Ruiz et al. 2013). Traditional optimization methods are either not able to handle the high dimensionality of NWP models or are impractical because they require too many model runs. On the other hand, sequential data assimilation methods treat model parameters as time-varying properties of the system. Those methods can be easily implemented in an existing data assimilation framework as they treat model parameters as extended state variables and are appropriate for estimating time-varying parameters such as leaf area index, surface roughness, and albedos. However, most parameters are formulated as constant properties of the system. Even if they are time varying, those model parameters and state variables do not vary on the same time scale. A two-stage filtering is needed to estimate model parameters and model state variables separately (Santitissadeekorn and Jones 2015; Vrugt et al. 2005). More recently, methods specially designed for parameter estimation of large complex system models like NWP have shown promise in improving the forecasting skill of NWP and climate models (Bellprat et al. 2012; Johnson et al. 2015; Neelin et al. 2010).

This paper demonstrates one such automatic model calibration platform for optimizing the parameters of NWP models. The keys of this model calibration platform are 1) to reduce the number of tunable parameters to a tractable level (e.g., from the current tens to hundreds to 15 or less) and 2) to develop an optimization algorithm that requires a relatively small number of model runs (only a few hundred at most). To implement these keys, a number of mathematical techniques can be employed, including 1) using a design-of-experiment (DoE) approach to judiciously sample model parameter sets within their physical variability ranges, and then using global sensitivity analysis (GSA) methods to

Table 1. The WRF Model parameterization schemes used in this study.

Physical process	Specific scheme
Surface layer	Monin–Obukhov scheme (Dudhia et al. 2005)
Cumulus	Kain–Fritsch (new Eta) scheme (Kain 2004)
Microphysics	WRF single-moment 6-class graupel scheme (Hong and Lim 2006)
Shortwave radiation	Dudhia scheme (Dudhia 1989)
Longwave radiation	Rapid Radiative Transfer Model scheme (Mlawer et al.1997)
Land surface	Unified Noah land surface model scheme (Chen and Dudhia 2001)
Planetary boundary layer	Yonsei University scheme (Hong et al. 2006)

identify the parameters that have the most impact on model forecasts; 2) instead of optimizing model parameters directly by running the NWP model repeatedly, constructing a surrogate model (also called statistical emulator, or metamodel) to represent the error response surface of the dynamical NWP model using a finite small number of model runs; and 3) using a multiobjective optimization approach to find the optimal parameters of the surrogate model and then using them to approximate the optimal parameters of the NWP model. In the next section, we provide a brief description of the platform. In the sections thereafter, we illustrate the usefulness of the automatic calibration platform through a case study involving 5-day forecasting of summer precipitation and surface air temperature in the greater Beijing area using the Weather and Research Forecasting (WRF) Model.

METHODS. A model calibration platform called Uncertainty Quantification Python Laboratory (UQ-PyL) has been developed, which has integrated different kinds of uncertainty quantification (UQ) methods, including various DoE, GSA, surrogate modeling, and optimization methods. It is written in Python language (PyL) and can run on all common operating systems such as Windows, Linux, and MacOS, with a graphical user interface for selecting and executing various commands. The different functions of UQ-PyL have been documented in Wang et al. (2016) and some specific UQ methods are described in several publications (Gan et al. 2014; Li et al. 2013; Wang et al. 2014; Gong et al. 2016). For example, how to use different GSA methods to reduce the dimensionality of complex dynamical models like the Common Land Model (CoLM) and the WRF Model has been demonstrated by Li et al. (2013), Di et al. (2015), and Quan et al. (2016). The idea behind parameter dimensionality reduction is to use sensitivity analysis to screen out the most important parameters that exert significant influence on model predictions (Guo et al. 2014). Once they are found, those parameters can then be optimized to maximize the model performance measures (Gong et al. 2015).

UQ-PyL also includes tools for constructing surrogate models and for conducting surrogate-modeling-based optimization. In Wang et al. (2014), an adaptive surrogate-modeling-based optimization (ASMO) method was described, which allows for effective and efficient searches of optimal parameters of large complex models using a low number of model runs. The goal of optimization is to find the minimum of an error response surface in the multiparameter

Table 2. The list of sensitive WRF Model parameters for precipitation and surface air temperature forecasts identified through sensitivity analysis.

Scheme	Parameter index	Parameter name	Default value	Parameter range	Parameter description
Cumulus (module_cu_kfeta.F)	P3	pd	1	[0.5, 2]	Multiplier for downdraft mass flux rate
	P4	pe	1	[0.5, 2]	Multiplier for entrainment mass flux rate
	P5	ph	150	[50, 350]	Starting height of downdraft above updraft source layer (hPa)
Microphysics (module_mp_wsm6.F)	P8	ice_stokes_fac	14,900	[8,000, 30,000]	Scaling factor applied to ice fall velocity (s^{-1})
	P10	dimax	5×10^{-4}	[3×10^{-4}, 8×10^{-4}]	Limited maximum value for the cloud ice diameter (m)
Shortwave radiation (module_ra_sw.F)	P12	cssca	1×10^{-5}	[5×10^{-6}, 2×10^{-5}]	Scattering tuning parameter ($m^2\ kg^{-1}$)
Land surface (module_sf_noahlsm.F)	P16	porsl	1	[0.5, 2]	Multiplier for the saturated soil water content
Planetary boundary layer (module_bl_ysu.F)	P20	Brcr_sb	0.25	[0.125, 0.5]	Critical Richardson number for boundary layer of land
	P21	pfac	2	[1, 3]	Profile shape exponent for calculating the momentum diffusivity coefficient

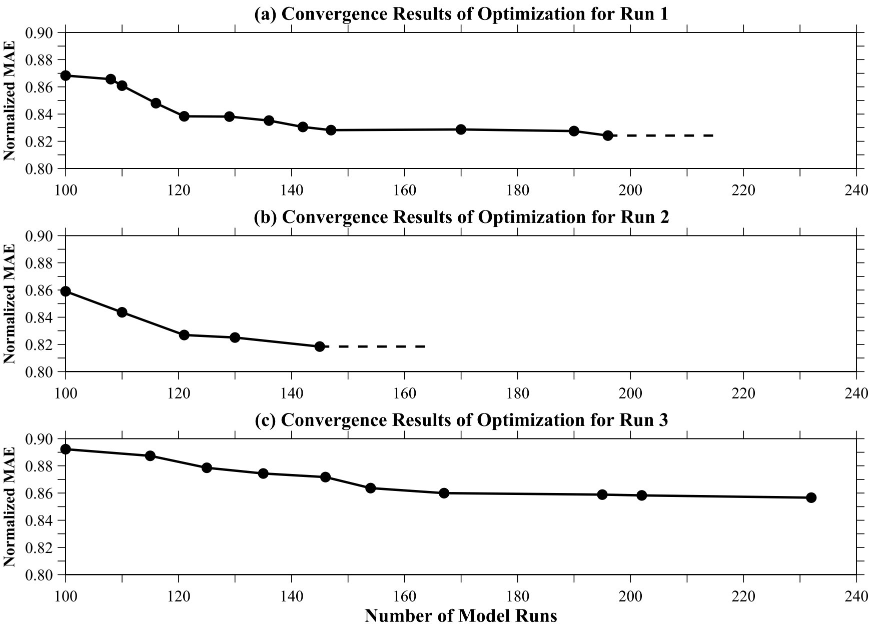

Fig. 2. The convergence results of three optimization runs: Normalized MAE values vs the number of model runs. The normalized MAE value for the control run (i.e., simulations using the default parameters) is equal to 1. If the normalized MAE is less than 1, it indicates that an improvement has been achieved over the control run.

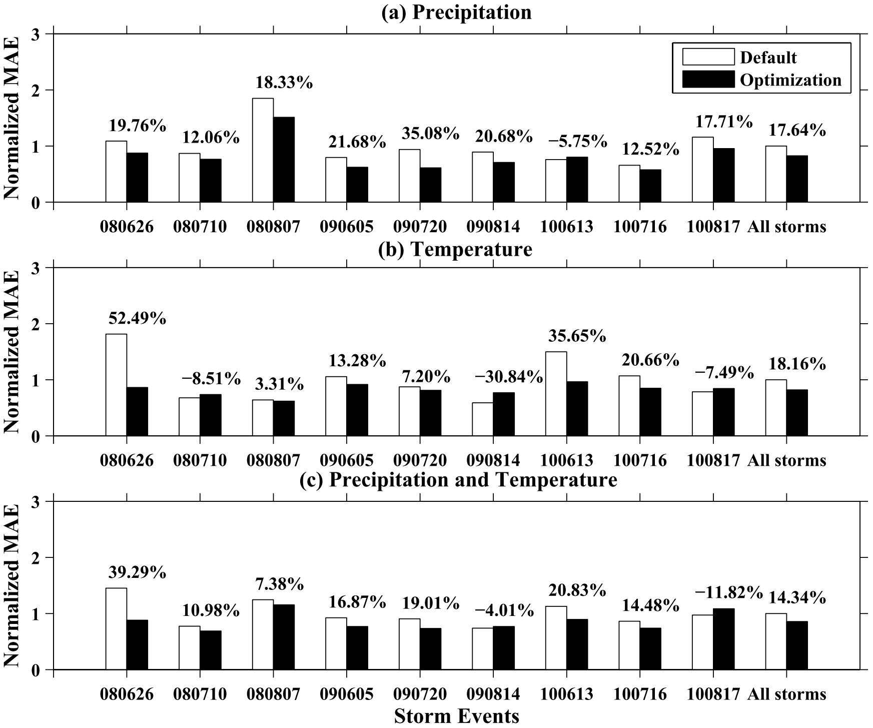

Fig. 3. Comparison of the objective function values of the three optimization runs against the objective function of the control run for nine individual calibration events as well as for all events combined. The numbers shown above the bars are the relative improvement of the optimized results vs the control run results.

space. ASMO is based on the premise that the optimal solution of a dynamical model can be approximated by the optimal solution of a surrogate model, which is constructed in two steps. The first step uses a DoE approach to create random parameter samples that are evenly distributed within the physical variability ranges of the parameters and then construct an error response surface using these parameter samples. The second step is to refine this response surface iteratively by using an adaptive sampling strategy that places more parameter samples in the promising parameter space based on information already gained on the existing response surface. Once this iterative process converges, the final response surface is treated as the surrogate model. The optimal solution of this surrogate model should approximate the optimal solution of the dynamical model.

NUMERICAL CASE STUDY.

We demonstrate how UQ-PyL can be used to calibrate the WRF Model through a case study. Our goal is to optimize the WRF Model parameters that are most important to the forecasting of precipitation P and surface air temperature T over the greater Beijing area during the summer monsoon. Nine 5-day precipitation events from the June–August (JJA) period between 2008 and 2010 are used for model calibration and

15 additional nonoverlapping 5-day precipitation events from 2005 to 2010 are chosen for validation of the optimization results (see Figs. ES1 and ES2). Those nine events were chosen for model calibration because they captured most of the heavy storm events over those years. The JJA 3-month period is chosen for calibration because over 70% of the annual precipitation in the area occurs during this period. Furthermore, storm events during this period are often associated with severe urban flooding that has resulted in tremendous economic and even human losses.

WRF Model, version 3.3, available from the WRF web portal (www2.mmm.ucar.edu/wrf/users/download/get_source.html), is used in the study. The setup for the WRF Model is described in Di et al. (2015) and the specific parameterization schemes used in this study are shown in Table 1. The WRF Model is run with a two-level nested domain setup over the greater Beijing area (Fig. 1), with the grid resolution of the outer domain (marked as "d01") at 27 km and the inner domain (marked as "d02") at 9 km. The model performance is evaluated over the d02 domain. The lateral and initial conditions needed to run the WRF Model are set using the NCEP Reanalysis data (Kistler et al. 2001). Based on the sensitivity analysis results of Di et al. (2015) and Quan et al. (2016), 9 parameters (see Table 2) selected out of a list of 23 tunable parameters (see the entire list of parameters in the supplemental materials; Table ES1; http://dx.doi.org/10.1175/BAMS-D-15-00104.2) were identified as important to *P* and *T* forecasting over the area. Subsequent surrogate modeling is therefore constructed using those nine parameters as independent variables. We employed the ASMO method to optimize those parameters. Four optimization runs were conducted with the ASMO method. The first two optimization runs were to minimize the normalized mean absolute errors (MAEs) of the *P* and *T* forecasts, respectively, while the third optimization run aimed to minimize the equally weighted normalized MAEs of *P* and *T* forecasts. The fourth optimization run was done using the ASMO method to maximum the weighted threat score (TS) of *P* forecasts for different storm categories.

The normalized MAE (NMAE), used in the first two optimization runs as the objective function (also called cost function), is computed based on the ratio of mean absolute difference between model prediction and observation over all grid points and forecasted time intervals using the given and default parameters; that is,

$$\text{minimize NMAE} = F(\theta) = \frac{f(\theta)}{f(\theta^*)}, \tag{1}$$

where $f(\cdot) = \left[\sum_{t=1}^{M}\sum_{i=1}^{N}\left|\text{sim}_i^t(\cdot) - \text{obs}_i^t\right|\right] \Big/ MN$, θ, and $\theta^* \in \Theta$ are the given and default parameter vectors, $\text{sim}_i^t(\theta)$ is the forecasted daily value at grid *i* and time *t* given θ, obs_i^t is the observation, *N* is the number of grid cells in domain d02, and *M* is the total number of forecasted time intervals. If $\theta = \theta^*$, $F(\theta) = 1$. If $F(\theta) < 1$, it implies that θ would produce better forecasts than θ^*. For optimization run 3, we used a weighted multiobjective function suggested by Liu et al. (2004) as we are concerned with minimizing the NMAE values of both *P* and *T* forecasts. The specific formulation of the objective function $F'(\theta)$ for this run is as follows:

$$\text{minimize } F'(\theta) = \sum_{j=1}^{2} w_j F_j(\theta), \tag{2}$$

where w_j is the weight for variable *j*, *j* = 1 denotes *P*, and *j* = 2 denotes *T*, respectively. We assigned equal weights to both *P* and *T* (i.e., $w_1 = w_2 = 0.5$).

For optimization run 4, we used TS [also known as critical success index (CSI)] as the objective function. We compuate TS, which measures the fraction of forecast events that are correctly predicted based on observations, as follows:

$$\text{TS} = \frac{\text{na}}{\text{na} + \text{nb} + \text{nc}}, \tag{3}$$

where na is the grid counts of hits (i.e., both forecast and observation fall in prescribed threshold ranges), nb is the grid counts of false alarms (i.e., the forecast falls in the threshold ranges, while observation does not), and nc is the grid counts of misses (i.e., the forecast falls outside the threshold ranges, while observation falls in). For a 6-h precipitation event [mm (6 h)$^{-1}$], the threshold values (mm) are set for six categories of precipitation events: light rain, moderate rain, heavy rain, storm, heavy storm, and severe storm are [0.1, 1), [1, 5), [5, 10), [10, 25), [10, 50), [50, ∞) mm, respectively. A higher TS score

Table 3. Comparison of the optimized objective function values for the first three optimization runs. The boldface numbers indicate the best objective function values for *P* and *T*.

Optimization run	Objective function value for *P* forecasts	Objective function value for *T* forecasts
1	**0.82359**	0.93526
2	0.90815	**0.81845**
3	0.87598	0.83729

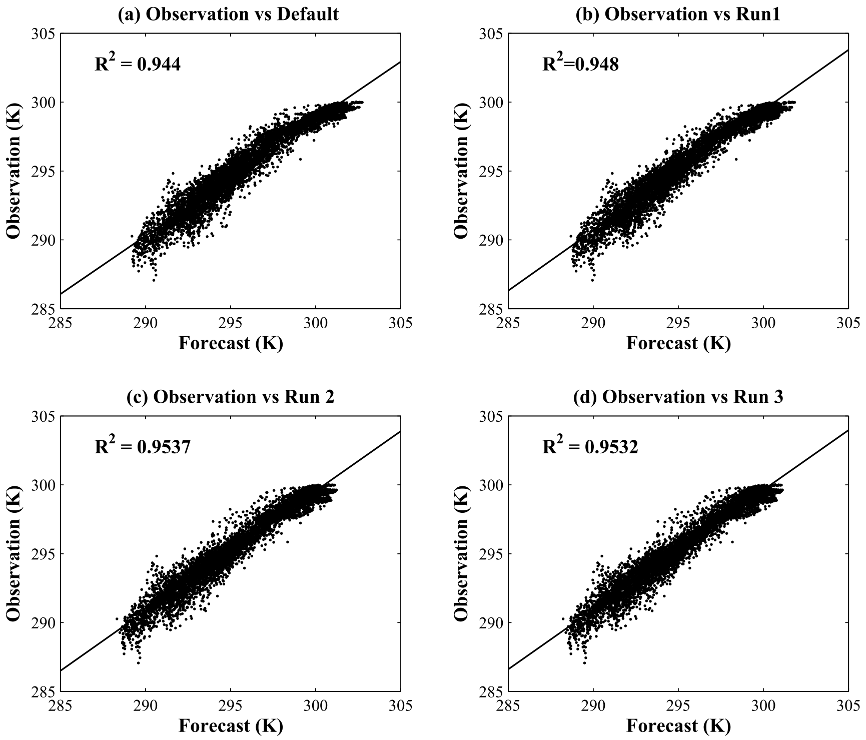

Fig. 4. The scatterplots of the 3-h surface air temperature for all grid points from the nine calibration events: (a) observation vs simulation using the default parameters, (b) observation vs simulation using optimized parameters from optimization run 1, (c) observation vs simulation using optimized parameters from optimization run 2, and (d) observation vs simulation using optimized parameters from optimization run 3.

indicates a better forecast, with a perfect score being 1. For the calibration events chosen for this case study, we only have the first four categories. In the optimization run, we used a weighted TS score, with a weight of 0.25 assigned to those four categories of storms.

Besides NMAE and TS scores, we also computed a combination score named *SAL* (Wernli et al. 2008) to evaluate the forecast performance, where *S*, *A*, and *L* stand for the forecast performance in terms of structure, amplitude, and location of the storm, respectively. A large value for *S* with a range of [–2, 2] implies the forecast storm area is too broad or too flat compared to observation; a small value means the forecast is too peaked and a value of 0 indicates being perfect. Amplitude *A*, with a range of [–2, 2], measures the average bias in precipitation amount, with 0 being perfect, a positive value indicating overforecast, and a negative value indicating underforecast. Location *L*, with a range of [0, 2], measures the displacement of forecast storm center from observed storm center, with 0 being perfect. The exact formulas for computing those scores are given in supplemental materials.

For all optimization runs, the size of the initial set of parameter samples used to construct the initial surrogate model was set to 100, based on the suggestion by Wang et al. (2014). According to the ASMO algorithm, from 45 to over 130 additional parameter sets were sampled adaptively to obtain the final surrogate model and optimal parameter set for the first three optimization runs (see Fig. 2). Figure 3 compares the optimized NMAE values against those corresponding to the control run (i.e., the run made using the default parameters) for the nine individual 5-day calibration events as well as for all calibration events combined. The results clearly indicate that optimization has resulted in significant improvement in the NMAE values. The improvement for all events combined ranges from 14.34% for the third optimization run to 18.16% for the second optimization run. The improvement is more evident when the NMAE of a single variable (e.g., *P* or *T*) is optimized than when the weighted NMAE of both *P* and *T* forecasts is optimized. From Table 3 we note that the optimized NMAE values for the individual variables yield their respective best values (i.e., optimization run 1 yields the best value for *P*, while optimization run 2 for *T*) and the optimized weighted NMAE value of both *P* and *T* corresponds to the compromised values (i.e., they are better than default but worse than the individually optimized values).

When examining the relative improvement for the individual calibration events, we see improvement in most events for all three optimization runs. However, there are a few events in which the NMAE values are not improved by optimization. But those events tend to correspond to the events whose NMAE values are relatively small compared to that of other events. Figures 4 and 5 show the scatterplots and the coefficients of determination R^2 between the 3-h *T* and *P* forecasts and corresponding observations for all grid points during the calibrated events using the default parameters and three sets of optimized parameters. For *T* forecasts, the R^2 values of all three optimization runs have improved over that of the default parameters,

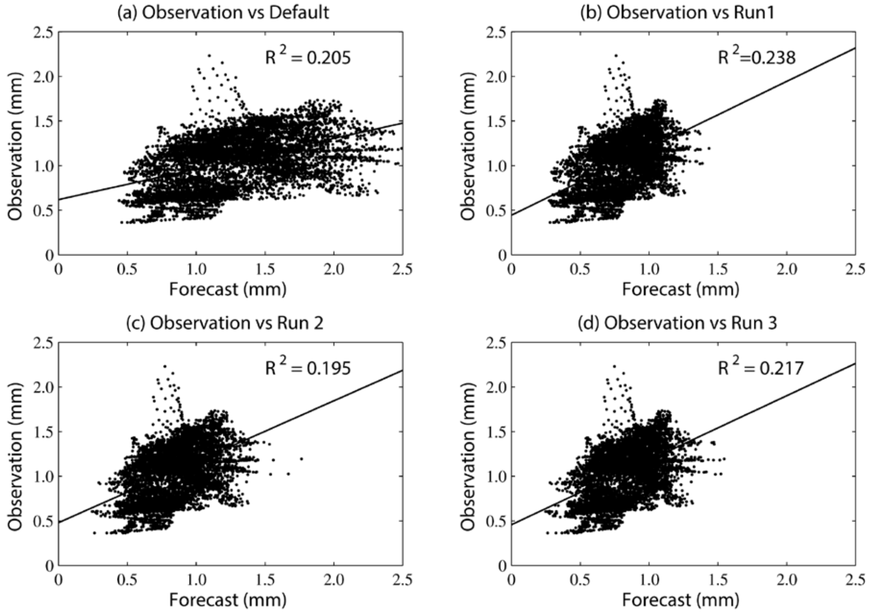

FIG. 5. The scatterplots of the 3-h precipitation for all grid points from the nine calibration events: (a) observation vs simulation using the default parameters, (b) observation vs simulation using optimized parameters from optimization run 1, (c) observation vs simulation using optimized parameters from optimization run 2, and (d) observation vs simulation using optimized parameters from optimization run 3.

with the second run (which optimizes *T* forecasts) having the highest R^2. For *P* forecasts, the R^2 values for optimization run 1 and run 3 have improved over that of the default parameters, but run 2 has degraded R^2 from that of default value. This tells us that optimizing *T* forecasts may not result in improved *P* forecasts.

We also examined the optimization results for different lead times and found that the improvement is consistent for all lead times. Figure 6 shows the relative improvement of the *P* forecasts for different lead times for all calibration events, which ranges from 10.87% for day 1 to 24% for day 2. We further examined if the optimal parameters from run 3 would lead to improved forecasts for other meteorological variables such as 2-m relative humidity, surface air pressure, 10-m wind speed, and downward surface shortwave radiation. Figure 7 confirmed that the optimal parameters for *P* and *T* forecasts have indeed improved the forecasts of those variables, with improvement rate of 1.71% for surface air pressure to 27% for wind speeds. Figure 8 presents the *SAL* score for the calibration events for run 3 and shows that the optimized parameters lead to better *SAL* scores compared to the default parameters. The improvement due to optimized parameters is the most obvious for structure *S* and amplitude *A*, implying the storm-area coverage and magnitude are better, and slight for storm location *L* prediction. The large value for *A* for the default parameters is consistent with the finding that *P* forecasts using the default parameters overestimate observed *P* (see Fig. ES3, which compares the daily averaged *P* forecasts against observations over the entire nine events for the optimized and default parameters).

We have computed the performance measures for the individual validation events as well as for all validation events combined. The improvement in NMAE values, as well as for other performance measures such as R^2, the SAL score, and the NMAE values for different lead times are similar to or slightly worse than those for the calibration events (see Figs. ES5–ES9 for detailed results).

Optimizing the WRF Model parameters using NMAE as the objective function has resulted in better performance according to a number of performance measures shown previously. However, it does not always lead to consistent improvement in the TS values (see Fig ES10, which shows the TS values of the *P* forecasts using the optimized parameters from optimization run 1). Since TS is a key performance measure emphasized by many operational meteorologists, we conducted an additional optimization run by using the weighted TS for different storm categories as the objective function. The optimization results indicate that the weighted TS was improved by 9.5% after 162 model runs (see Fig. ES11, which shows the convergence of optimization run 4). Figure 9 shows the TS scores for individual calibration events as well as for all calibration events combined for the four categories of storms, with the TS scores for the combined events varying from −1.09% for light rain to 23.88% for heavy rain. The decreased TS score for light rain is because a weighted TS is used as the objective function. Even though the overall TS would improve, a TS for an individual category may not. This actually indicated a problem when a single objective function is used for model calibration. Many studies have suggested that a truly multiobjective model calibration approach can be useful, in which Pareto optimal parameter sets are identified (Gong et al. 2016). In those

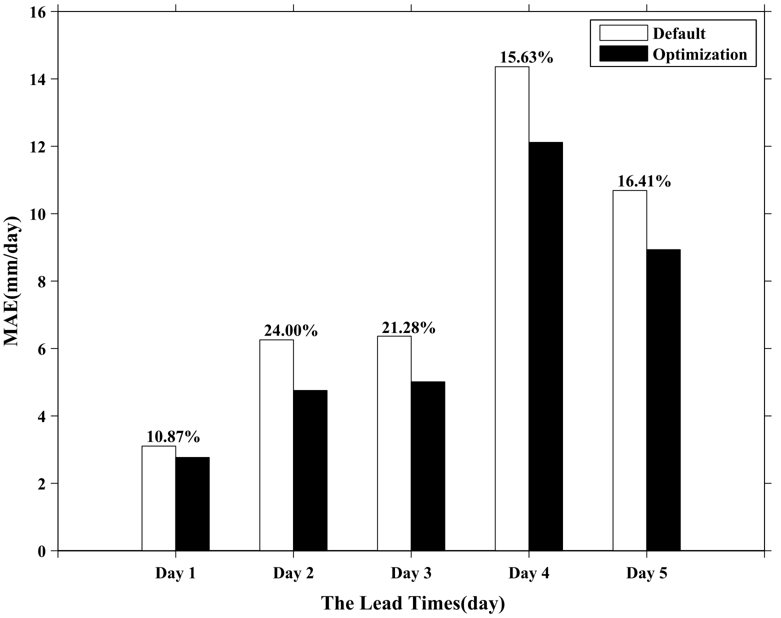

Fig. 6. Comparison of the MAE values for precipitation based on simulations using the default parameters and using the optimized parameters for different lead times for all calibration events combined.

Pareto optimal sets, some parameter sets correspond to optimal forecasting performance for light rain events, others being optimal for moderate or heavy rain events. None of the objective functions in the Pareto optimal sets can be improved in value without degrading some of the other objective values (Miettinen 1999). The optimal parameters obtained using TS as the objective function were validated using independent validation datasets. We found that TS for all validation events combined have been improved, even though TS for individual storms may not (see Fig. ES12).

We examined the differences in the optimized parameter values for all optimization runs and compared them against the default values (Fig. 10). Direct correspondence between the parameter values and model performance measure is not obvious, because *P* and *T* forecasts are the aggregated results of complicated, highly nonlinear interactions between different physical processes such as ascent or descent of moist air, turbulent exchanges of water and energy fluxes between land surface and atmosphere, and horizontal advection of momentum, mass, and energy. From Fig. 8 (also Figs. ES3 and ES4), there is an apparent overestimation of both *P* and *T* over the greater Beijing area when default parameters are used. The changes in most parameters tend to show the effect of depressing *P* and lowering *T*. For example, a large value for P12 (the scattering tuning parameter) means a higher scattering of solar radiation, leading to reduced shortwave radiation reaching the surface and thus decreasing evaporation from the ground, and ultimately depressed *P* and *T*. The changes in P3 (the multiplier for downdraft mass flux rate) and P5 (the starting height of downdraft above the updraft source layer) have similar effects. When

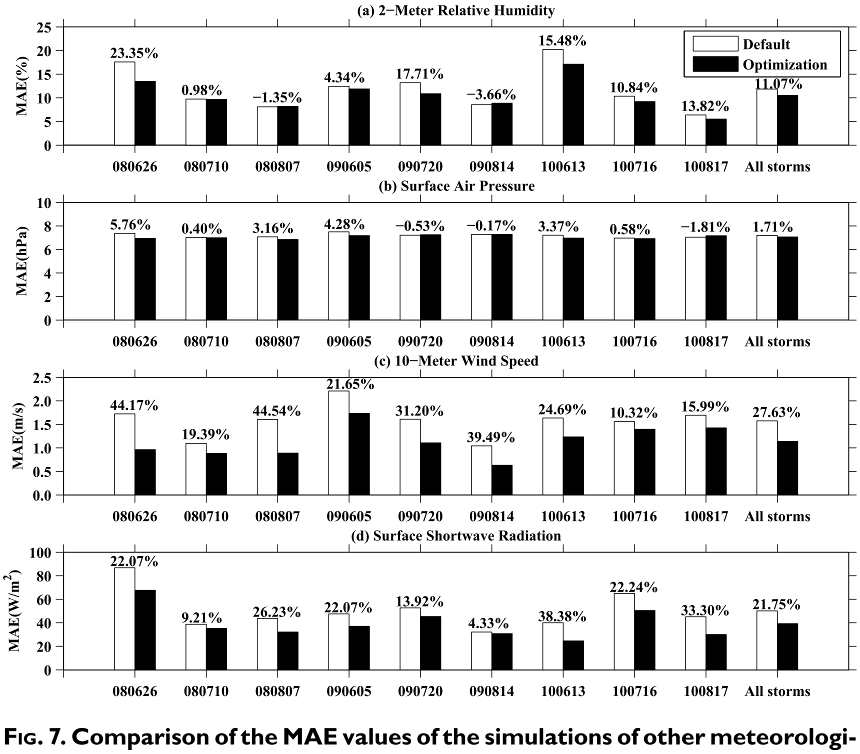

Fig. 7. Comparison of the MAE values of the simulations of other meteorological variables using the default parameters and using the optimal parameters for the nine calibration events and for all calibration events combined: (a) 2-m relative humidity, (b) surface air pressure, (c) 10-m wind speed, and (d) surface shortwave radiation.

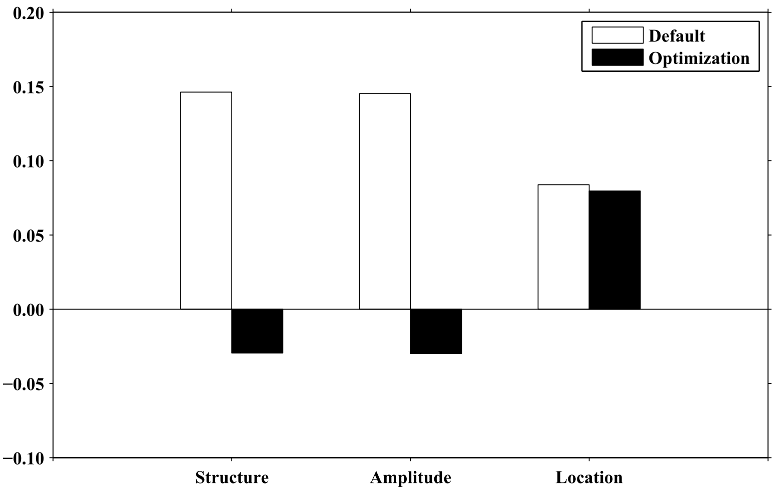

Fig. 8. Comparison of the *SAL* precipitation forecasting skill score for the nine calibration events combined. For all *SAL* components, a value of 0.0 indicates a perfect score.

their values increase, the downdraft flux becomes stronger, suppressing further development of updraft convection and depressing precipitation. Meanwhile, more evaporation from condensed water occurs during the downdraft, cooling the atmosphere temperature, thus reducing *T*. For P16 (the multiplier for the saturated soil water content), the optimized values for *P* and *T* are conflicting. When P16 value increases, soil moisture content increases and surface evapotranspiration is enhanced, thus inducing stronger *P*. On the other hand, when P16 value decreases, the incidence of *P* becomes weak and higher *T* results. Parameter P21 (the profile shape exponent for calculating momentum diffusivity coefficient) reflects the mixing intensity of turbulent eddies in planetary boundary layer. When its value decreases, turbulence diffusivity intensity is weakened, making the upward transfer of heat and water vapor from the ground surface slow down. So the formation of convection is more difficult, and *P* is thus reduced. The slower thermal eddy diffusivity also restrains the increase of *T*. For parameters P4 (multiplier of entrainment mass flux rate) and P8 (scaling factor applied to ice fall velocity), lower values for them lead to less favorable conditions for formation of rain and thus lead to reduced *P*. A larger value for P10 (the maximum value for the cloud ice diameter) has similar effect as the lower P8. When parameter P20 (the critical Richardson number for boundary layer) decreases, the planetary boundary layer height is depressed and thermal eddy diffusivity is decreased, leading to lower *T*. Of course, it is impossible to explain all the changes in parameter values completely because of nonlinear interactions among them.

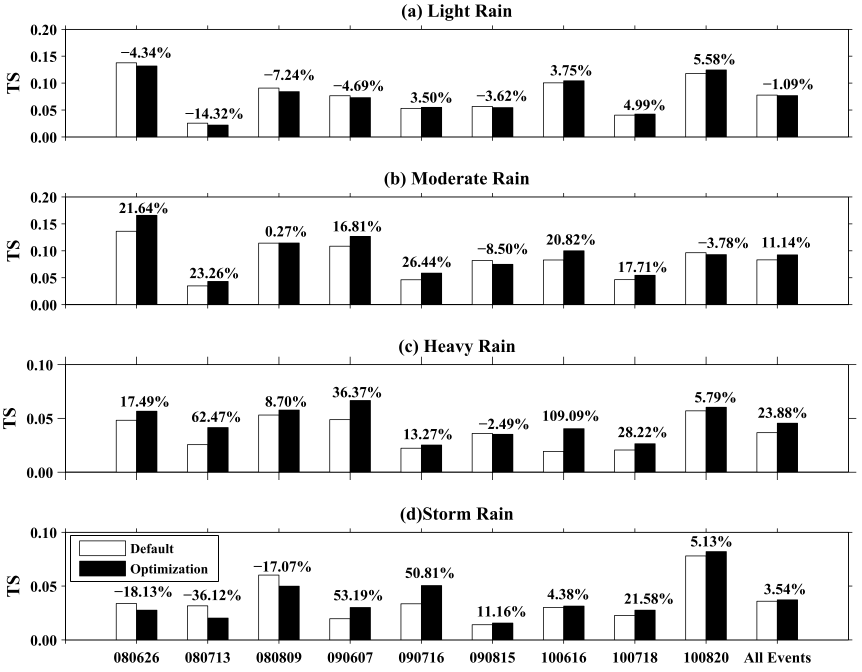

Fig. 9. Comparison of the TS values of the precipitation simulations obtained using the default parameters and using the optimized parameters from optimization run 4 for the nine calibration events and all calibration events combined.

CONSIDERATIONS FOR PRACTICAL APPLICATIONS.

We have demonstrated the potential of automatic model calibration as a new way to improve numerical weather forecasting. This study has been intended to provide practical guidance to operational NWP modelers. In practice, several considerations must be taken before using automatic model calibration methods. First, the model setup in this study is not exactly as in the operational settings for the greater Beijing area by the Beijing

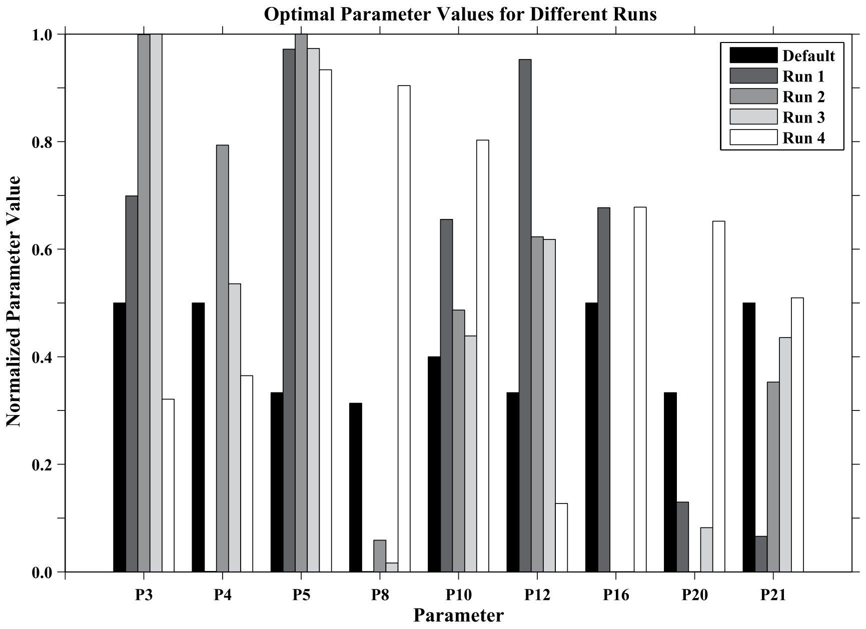

Fig. 10. Comparison of optimized parameter values against the default parameter values for the four optimization runs. The horizontal axis denotes different parameters and vertical axis denotes parameter values, with zero corresponding to the lower bound and 1 corresponding to the upper bound of the parameters.

Institute of Urban Meteorology Research, which runs the WRF Model over a larger domain with three-level nested grids and the inner grid resolution at 1 km, compared to a two-level nested domain and a 9-km resolution for the inner domain in this study. The operational model is also initialized with a data assimilation system that provides more realistic initial conditions, which should lead to further improvement because of improved initial conditions for the model. If the optimization methodology presented here is applied to the operational setting, we need to recalibrate the WRF Model and the final optimal parameter values may differ from the ones we obtained in this study. Another important note is that the tunable parameters we selected for this study may not include all impactful parameters on the forecasts, owing to the fact we are not totally familiar with all of the schemes in the WRF Model. Any WRF modeler who wants to optimize the WRF Model parameters should make full use of her knowledge about the model in identifying the tunable parameters.

Second, we choose nine summer storm events over a 3-yr period to ensure that the optimized model parameters are reasonably robust. We found that the improvement in calibration events is consistent when we computed a number of performance measures (i.e., NMAE, TS, R^2, the SAL scores, and NMAE for different lead times). More interestingly, the performance measures for other meteorological variables are also improved when P and T forecasts are optimized. The validation results confirm that the optimized parameters have also resulted in improved performance measures for the 16 individual validation events as well as for all validation events combined. When conducting model calibration for specific applications, one always needs to perform those validation studies to ensure the effectiveness of the optimization results. Further, even though the parameters obtained in this study were found to be optimal for forecasting summer storms, they may not be optimal for other types of storm events such as the winter storms. It may be arguable whether model calibration based on only nine calibration events is sufficient in practice, especially for the rare events. A more comprehensive study using a large sample of events may be necessary in the future to assess the impact of sampling error on optimal parameter estimates.

Third, the performance metrics used in calibration are very important. In this study we used NMAE and TS as the objective function and they resulted in different model parameters. We also used a number of performance measures such as the scatterplots, R^2, and the *SAL* scores to verify the forecasts. In practice, the evaluation of model performance is multifaceted and the choice of objective functions must reflect the values of the forecast practitioner. It is possible that the optimal parameters for one metric may conflict with that for another metric. One should consider a multiobjective approach that weighs different objective functions subjectively to reflect one's preference for a particular performance metric. Another approach may be used to search for a Pareto optimal set, in which all parameter sets are noninferior compared to another set according to at least one performance metric. With this set of Pareto optimal parameters, one can create an ensemble of forecasts using different parameters.

The optimization approach presented here is different from the sequential data assimilation approaches, which update model parameters as new data become available. The approach is also more efficient and effective than conventional optimization approaches (e.g., downhill simplex, simulated annealing, and genetic algorithm) as we relied on techniques such as parameter screening and surrogate modeling to achieve computational saving. We hope this study will stimulate more development in using automatic model calibration methods to optimize NWP model parameters.

ACKNOWLEDGMENTS. This research was partially supported by the Natural Science Foundation of China (41375139), the Ministry of Science and Technology of China (2015DFA20870),and the Beijing Municipal Science and Technology Commission (Z141100003614052 and Z151100002115045). We gratefully acknowledge the insightful comments from the editor and three anonymous reviewers whose advices have greatly enhanced the readability of this paper.

REFERENCES

Aksoy, A., 2015: Parameter estimation. *Encyclopedia of Atmospheric Sciences*, 2nd ed. G. R. North, J. Pyle, and F. Zhang, Eds., Academic Press, 181–186, doi:10.1016/B978-0-12-382225-3.00494-1.

Bellprat, O., S. Kotlarski, D. Lüthi, and C. Schär, 2012: Objective calibration of regional climate models. *J. Geophys. Res.*, **117**, D23115, doi:10.1029/2012JD018262.

Bennett, A. F., B. S. Chua, and L. M. Leslie, 1996: Generalized inversion of a global numerical weather prediction model. *Meteor. Atmos. Phys.*, **60**, 165–178, doi:10.1007/BF01029793.

Chen, F., and J. Dudhia, 2001: Coupling an advanced land surface–hydrology model with the Penn State–NCAR MM5 modeling system. Part I: Model implementation and sensitivity. *Mon. Wea. Rev.*, **129**, 569–585, doi:10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2.

Di, Z., and Coauthors, 2015: Assessing WRF model parameter sensitivity: A case study with 5-day summer precipitation forecasting in the greater Beijing area. *Geophys. Res. Lett.*, **42**, 579–587, doi:10.1002/2014GL061623.

Duan, Q., and Coauthors, 2006: Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results of the second and third workshops. *J. Hydrol.*, **320**, 3–17, doi:10.1016/j.jhydrol.2005.07.031.

Duane, G. S., and J. P. Hacker, 2007: Automatic parameter estimation in a mesoscale model without ensembles. *Nonlinear Time Series Analysis in Geophysics*, R. V. Donner and S. M. Barbosa, Eds., Lecture Notes in Earth Sciences, Vol. 112, Springer, 81–95.

Dudhia, J., 1989: Numerical study of convection observed during the winter monsoon experiment using a mesoscale two-dimensional model. *J. Atmos. Sci.*, **46**, 3077–3107, doi:10.1175/1520-0469(1989)046<3077:NSOCOD>2.0.CO;2.

Dudhia, J., D. Gill, K. Manning, W. Wang, and C. Bruyere, 2005: PSU/NCAR Mesoscale Modeling System tutorial class notes and user's guide: MM5 Modeling System version 3. NCAR Rep., 390 pp.

Evensen, G., D. P. Dee, and J. Schröter, 1998: Parameter estimation in dynamical models. *Ocean Modeling and Parameterization*, E.O. Chassignet and J. Verron, Eds., NATO Science Series, Vol. 516, Springer, 373–398.

Gan, Y., and Coauthors, 2014: A comprehensive evaluation of various sensitivity analysis methods: A case study with a hydrological model. *Environ. Modell. Software*, **51**, 269–285, doi:10.1016/j.envsoft.2013.09.031.

Gong, J., G. Wahba, D. R. Johnson, and J. Tribbia, 1998: Adaptive tuning of numerical weather prediction models: Simultaneous estimation of weighting, smoothing, and physical parameters. *Mon. Wea. Rev.*, **126**, 210–231, doi:10.1175/1520-0493(1998)126<0210:ATONWP>2.0.CO;2.

Gong, W., Q. Duan, J. Li, C. Wang, Z. Di, Y. Dai, A. Ye, and C. Miao, 2015: Multi-objective parameter optimization of common land model using adaptive surrogate modelling. *Hydrol. Earth Syst. Sci.*, **19**, 2409–2425, doi:10.5194/hess-19-2409-2015.

——, ——, ——, and Y. Dai, 2016: Multiobjective adaptive surrogate modeling-based optimization for parameter estimation of large, complex geophysical models. *Water Resour. Res.*, **52**, 1984–2008, doi:10.1002/2015WR018230.

Guo, Z., and Coauthors, 2014: A sensitivity analysis of cloud properties to CLUBB parameters in the single-column Community Atmosphere Model (SCAM5). *J. Adv. Model. Earth Syst.*, **6**, 829–858, doi:10.1002/2014MS000315.

Hong, S. Y., and J. J. Lim, 2006: The WRF single moment 6 class microphysics scheme (WSM6). *J. Korean Meteor. Soc.*, **42**, 129–151, doi:10.1155/2010/707253.

——, Y. Noh, and J. Dudhia, 2006: A new vertical diffusion package with an explicit treatment of entrainment processes. *Mon. Wea. Rev.*, **134**, 2318–2341, doi:10.1175/MWR3199.1.

Ihshaish, H., A. Cortes, and M. A. Senar, 2012: Parallel multi-level genetic ensemble for numerical weather prediction enhancement. *Proc. Comput. Sci.*, **9**, 276–285, doi:10.1016/j.procs.2012.04.029.

Jackson, C., M. Sen, and P. Stoffa, 2004: An efficient stochastic Bayesian approach to optimal parameter and uncertainty estimation for climate model predictions. *J. Climate*, **17**, 2828–2841, doi:10.1175/1520-0442(2004)017<2828:AESBAT>2.0.CO;2.

Johnson, J. S., Z. Cui, L. A. Lee, J. P. Gosling, A. M. Blyth, and K. S. Carslaw, 2015: Evaluating uncertainty in convective cloud microphysics using statistical emulation. *J. Adv. Model. Earth Syst.*, **7**, 162–187, doi:10.1002/2014MS000383.

Kain, J. S., 2004: The Kain–Fritsch convective parameterization: An update. *J. Appl. Meteor.*, **43**, 170–181, doi:10.1175/1520-0450(2004)043<0170:TKCPAU>2.0.CO;2.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation and Predictability.* Cambridge University Press, 368 pp.

Kistler, R., and Coauthors, 2001: The NCEP–NCAR 50-Year Reanalysis: Monthly means CD-ROM and documentation. *Bull. Amer. Meteor. Soc.*, **82**, 247–267, doi:10.1175/1520-0477(2001)082<0247:TNNYRM>2.3.CO;2.

Krishnamurti, T. N., C. M. Kishtawal, T. LaRow, D. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran, 1999: Improved skill of weather and seasonal climate forecasts from multimodel superensemble. *Science*, **285**, 1548–1550, doi:10.1126/science.285.5433.1548.

Li, J., and Coauthors, 2013: Assessing parameter importance of the Common Land Model based on qualitative and quantitative sensitivity analysis. *Hydrol. Earth Syst. Sci.*, **17**, 3279–3293, doi:10.5194/hess-17-3279-2013.

Liu, Y., H. V. Gupta, S. Sorooshian, L. A. Bastidas, and W. J. Shuttleworth, 2004: Exploring parameter sensitivities of the land surface using a locally coupled land-atmosphere model. *J. Geophys. Res.*, **109**, D21101, doi:10.1029/2004JD004730.

Miettinen, K., 1999: *Nonlinear Multiobjective Optimization.* International Series in Operations Research & Management Science, Vol. 12, Springer, 298 pp., doi:10.1007/978-1-4615-5563-6.

Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res.*, **102**, 16 663–16 682, doi:10.1029/97JD00237.

Mu, M., W. Duan, and J. Wang, 2002: The predictability problems in numerical weather and climate prediction. *Adv. Atmos. Sci.*, **19**, 191–204, doi:10.1007/s00376-002-0016-x.

Neelin, J. D., A. Bracco, H. Luo, J. C. McWilliams, and J. E. Meyerson, 2010: Considerations for parameter optimization and sensitivity in climate models. *Proc. Natl. Acad. Sci. USA*, **107**, 21 349–21 354, doi:10.1073/pnas.1015473107.

Quan, J., and Coauthors, 2016: An evaluation of parametric sensitivities of different meteorological variables simulated by the WRF model. *Quart. J. Roy. Meteor. Soc.*, **142**, 2925–2934, doi:10.1002/qj.2885.

Ruiz, J., M. Pulido, and T. Miyoshi, 2013: Estimating model parameters with ensemble-based data assimilation: A review. *J. Meteor. Soc. Japan*, **91**, 79–99, doi:10.2151/jmsj.2013-201.

Santitissadeekorn, N., and C. Jones, 2015: Two-stage filtering for joint state-parameter estimation. *Mon. Wea. Rev.*, **143**, 2028–2042, doi:10.1175/MWR-D-14-00176.1.

Severijns, C. A., and W. Hazeleger, 2005: Optimizing parameters in an atmospheric general circulation model. *J. Climate*, **18**, 3527–3535, doi:10.1175/JCLI3430.1.

Stensrud, D. J., 2007: *Parameterization Schemes: Keys to Understanding Numerical Weather Prediction Models.* Cambridge University Press, 480 pp.

Vrugt, J. A., C. G. H. Diks, H. V. Gupta, W. Bouten, and J. M. Verstraten, 2005: Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation. *Water Resour. Res.*, **41**, W01017, doi:10.1029/2004WR003059.

Wang, C., Q. Duan, W. Gong, A. Ye, Z. Di, and C. Miao, 2014: An evaluation of adaptive surrogate modeling based optimization with two benchmark problems. *Environ. Modell. Software*, **60**, 167–179, doi:10.1016/j.envsoft.2014.05.026.

——, ——, C. H. Tong, and W. Gong, 2016: A GUI platform for uncertainty quantification of complex dynamical models. *Environ. Modell. Software*, **76**, 1–12, doi:10.1016/j.envsoft.2015.11.004.

Wernli, H., M. Paulat, M. Hagen, and C. Frei, 2008: SAL—A novel quality measure for the verification of quantitative precipitation forecasts. *Mon. Wea. Rev.*, **136**, 4470–4487, doi:10.1175/2008MWR2415.1.

Yu, X., S. K. Park, and Y. H. Lee, 2013: Parameter estimation using an evolutionary algorithm for QPF in a tropical cyclone. *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications*, Vol. II, S. K. Park and X. Liang, Eds., Springer, 707–715, doi:10.1007/978-3-642-35088-7_27.