Use of a Genetic Algorithm to Optimize a Numerical Weather Prediction System

P. L. HOUTEKAMER,^a BIN HE,^a DOMINIK JACQUES,^a RON McTAGGART-COWAN,^a LEO SEPAROVIC,^a
PAUL A. VAILLANCOURT,^a AYRTON ZADRA,^a AND XINGXIU DENG^b

^a *Meteorology Research Division, Environment and Climate Change Canada, Dorval, Québec, Canada*

^b *Meteorological Service of Canada, Environment and Climate Change Canada, Dorval, Québec, Canada*

(Manuscript received 21 July 2020, in final form 3 December 2020)

ABSTRACT: An important step in an ensemble Kalman filter (EnKF) algorithm is the integration of an ensemble of short-range forecasts with a numerical weather prediction (NWP) model. A multiphysics approach is used in the Canadian global EnKF system. This paper explores whether the many integrations with different versions of the model physics can be used to obtain more accurate and more reliable probability distributions for the model parameters. Some model parameters have a continuous range of possible values. Other parameters are categorical and act as switches between different parameterizations. In an evolutionary algorithm, the member configurations that contribute most to the quality of the ensemble are duplicated, while adding a small perturbation, at the expense of configurations that perform poorly. The evolutionary algorithm is being used in the migration of the EnKF to a new version of the Canadian NWP model with upgraded physics. The quality of configurations is measured with both a deterministic and an ensemble score, using the observations assimilated in the EnKF system. When using the ensemble score in the evaluation, the algorithm is shown to be able to converge to non-Gaussian distributions. However, for several model parameters, there is not enough information to arrive at improved distributions. The optimized system features slight reductions in biases for radiance measurements that are sensitive to humidity. Modest improvements are also seen in medium-range ensemble forecasts.

KEYWORDS: Ensembles; Genetic algorithms/programming

1. Introduction

Numerical weather forecasts increasingly diverge from the observed atmospheric state for longer prediction lead times. Among the reasons for the divergence are inaccurate initial conditions and the use of a forecast model that cannot exactly mimic the atmospheric processes and dynamics. With an ensemble prediction system, one samples the dominant sources of error to subsequently arrive at a prediction of the distribution of the forecast error (Houtekamer et al. 1996). The sampling of error, by means of using different configurations of the model physics, has been part of the Canadian Global Ensemble Prediction System (GEPS) since its first operational use. Unfortunately, after over 20 years of development (Houtekamer and Lefaire 1997; Houtekamer 2011), we have yet to arrive at an effective procedure for obtaining optimal model configurations that can evolve with the ongoing development of our center's NWP system. A set of major changes to the model physical parameterizations (McTaggart-Cowan et al. 2019b) provided the impetus for the current study, in which a genetic algorithm is used to estimate model parameters and their uncertainty.

An interesting possibility to reduce uncertainty in model parameters is to use ensemble data assimilation techniques (Grell and Dévényi 2002; Ruiz et al. 2013; Schirber et al. 2013). The simultaneous estimation of the model state and six model

parameters is discussed by Aksoy et al. (2006) in the context of a two-dimensional sea-breeze model. It may, however, be advantageous to disconnect the estimation of a large number of quickly changing local state variables, the evolution of which needs to be tracked, from the estimation of a small number of model parameters. The latter have more universal validity and should not be changing on short time scales.

With the ensemble prediction and parameter estimation system (EPPES; Järvinen et al. 2012), it is proposed to couple an operational ensemble prediction system with the online estimation of model parameters. The general idea is to replace model configurations that perform poorly with configurations that perform well. Such replacement is repeated over a period of time, after which the parameters will have evolved to more optimal values. In a companion paper (Laine et al. 2012), the method is demonstrated within a low-order experimental environment.

In the Canadian operational system, a global ensemble Kalman filter (EnKF), with $N_{\text{ens}} = 256$ members and a 6-h cycling interval, is used to provide the initial conditions for a 20-member global medium-range ensemble prediction system that is run twice daily. Multiple physical parameterizations are used in both systems. When used for parameter estimation, the EnKF could, in principle, provide up to 1024 evaluations of proposed parameter values per day. In contrast, the 20-member ensemble prediction system (EPS) can provide only up to 40 evaluations per day. As the EnKF environment can provide an order of magnitude more evaluations than the medium-range EPS, it is used here in an EPPES.

In our implementation, each ensemble member is used to evaluate its specific set of parameters. Based on an evaluation function, sets of parameter values can either be removed or

Denotes content that is immediately available upon publication as open access.

Corresponding author: P. L. Houtekamer, peter.houtekamer@canada.ca

duplicated and perturbed. Here, the addition of a small perturbation serves to maintain a realistic and sufficient diversity in the ensemble and thus avoid the premature collapse of the ensemble. Such a selection among proposed values is common in genetic and particle filter algorithms (Gordon et al. 1993; Snyder et al. 2008). It permits arriving at multimodal non-Gaussian distributions and can be used with categorical parameters such as switches in model physical parameterizations.

Here we are not aiming at a particle-filter-like optimal algorithm that rapidly converges to the unique best parameter values but rather, in an effort to represent the inherent complexity of model physical parameterizations (Ruiz et al. 2013), at a reliable algorithm that provides broader distributions that continue to be appropriate for a diverse set of atmospheric conditions. In line with this thinking, we refer to our algorithm as a genetic algorithm.

Díaz-Isaac et al. (2019) use the flatness of rank histograms as the main criterion to arrive at a subensemble with optimal reliability. They also recommend including the statistical resolution in the optimization procedure. Following their guidance, in this study the continuous ranked probability score (CRPS; Wilks 2006), which measures both the statistical resolution and the statistical reliability of the ensemble, is tested for use in the evaluation step along with a traditional quadratic cost function.

Section 2 describes the scores and the algorithm that are used for the optimization. Section 3 describes the parameterization of uncertainty that provides the context for the optimization work. In section 4, results of an optimization experiment are given for a 20-day period. We end in section 5 with the conclusions.

2. The genetic algorithm

We use a genetic algorithm to identify sets of parameter values that optimize the ensemble performance. Section 2a describes two functions that can be used to evaluate quality. The probabilistic CRPS computes quality using all available members simultaneously. In contrast, a traditional cost function provides a score to evaluate each member individually. Section 2b describes the optimization algorithm that, with minor modifications, can be used with either of the above two types of scores.

a. The measures for quality

There are two clear options for the selection of members to be replaced by the genetic algorithm, both of which attempt to identify members that are performing poorly. The first is an ensemble-based approach, in which each member is assessed based on its contribution to the cost function for the overall ensemble. A natural score for this is the CRPS. For an observation y_k and a verifying ensemble $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_{N_{\text{ens}}}]$, the CRPS can be computed as (Gneiting and Raftery 2007; Zamo and Naveau 2018):

$$\text{CRPS}(\mathbf{X}, y_k) = \frac{1}{N_{\text{ens}}} \sum_{i=1}^{N_{\text{ens}}} |\mathcal{H}_k \mathbf{x}_i - y_k| - \frac{1}{2N_{\text{ens}}^2} \sum_{i,j=1}^{N_{\text{ens}}} |\mathcal{H}_k \mathbf{x}_i - \mathcal{H}_k \mathbf{x}_j|. \quad (1)$$

Here, the possibly nonlinear forward operator $\mathcal{H}_k$ is used to compute the model-based equivalent of the observation y_k . Thus, Eq. (1) assigns a single score to the difference of the probability distribution function of the ensemble of states $\mathbf{X}$ and the observation y_k . The score has the units of the observation y_k . The values of $\mathcal{H}_k \mathbf{x}_i$, $i = 1, \dots, N_{\text{ens}}$, $k = 1, \dots, N_{\text{obs}}$ are computed by the atmospheric EnKF and stored for subsequent use in the genetic algorithm. For more robust results, we use the observations y_k , $k = 1, \dots, N_{\text{obs}}$ used by the EnKF during 1 day rather than from a single 6-h window and the corresponding model states $\mathbf{x}_i$, $i = 1, \dots, N_{\text{ens}}$ consist of the four consecutive 6-h trajectories used by the EnKF during that day. The estimated standard deviation σ_k of the observational error is used for normalization:

$$\text{CRPS}(\mathbf{X}) = \frac{1}{N_{\text{obs}}} \sum_{k=1}^{N_{\text{obs}}} \frac{\text{CRPS}(\mathbf{X}, y_k)}{\sigma_k}. \quad (2)$$

With Eq. (2), we have a measure for the overall quality of the ensemble of background trajectories that is available from one day of cycling with the EnKF. To obtain a measure for the quality of the N_{ens} individual ensemble members, the score is computed N_{ens} times leaving out one member each time. Thus, to assign a cost $\mathcal{J}_i$ to member i , we compute the following:

$$\mathbf{X}^{-i} = [\mathbf{x}_1, \dots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \dots, \mathbf{x}_{N_{\text{ens}}}], \quad (3)$$

$$\mathcal{J}_i = \text{CRPS}(\mathbf{X}) - \text{CRPS}(\mathbf{X}^{-i}). \quad (4)$$

The second option, for the selection of members, relies on the deterministic performance of each member individually and is more consistent with the Bayesian framework of particle filters in which members are penalized for poor individual performance. In this spirit, we compute the score:

$$\mathcal{J}_1(\mathbf{x}_i) = \frac{1}{N_{\text{obs}}} \sum_{k=1}^{N_{\text{obs}}} \frac{(\mathcal{H}_k \mathbf{x}_i - y_k)^2}{\sigma_k^2}. \quad (5)$$

Note that this is identical to the observational component of the cost function that is commonly used in variational data assimilation.

The choice between these two penalty formulations will have a direct impact on the results of the optimization procedure because they import different concepts of value to the genetic algorithm. In the ensemble-based analysis, the members that are rejected are those that are most damaging to the ensemble: they create an imbalance between spread and skill as depicted by an increased CRPS. In the individually based analysis, the rejection is entirely dependent on the difference between the member and the set of verifying observations irrespective of the impact that the member has on the ensemble spread. The ensemble-based approach therefore leads the genetic algorithm to a set of parameters that yields a well-balanced ensemble, while the individually based analysis guides the process toward the parameters that minimize forecast error in a deterministic sense. We note that, in the context of an evolutionary algorithm, it would be straightforward to replace the L2 norm of Eq. (5) with an L1 norm. Whereas

this would still not directly measure ensemble quality, results would likely be more similar to those obtained using the CRPS.

b. The optimization algorithm

Here we first provide the optimization algorithm for use with Eq. (4), and subsequently we discuss the modification required when using Eq. (5).

Step 1: Initial selection of parameters.

To start the experiment, select $N_{\text{ens}} = 256$ sets of N_{par} model parameters. Each parameter set will be used by one member of the EnKF that serves to analyze the atmospheric state.

Step 2: Test phase.

During the test phase, a number of 6-h assimilation steps ($N_{\text{analysis}} = 4$) is performed with the EnKF. The test phase thus covers a 24-h period and is motivated by the existence of a daily cycle in both weather phenomena and in the observational network. Using a shorter optimization period could cause oscillatory behavior of parameter estimates. Note that the EnKF serves only to analyze the atmospheric state every 6 h. It does not provide estimates of model parameters. As is standard in our stochastic EnKF system, different ensemble members assimilate differently perturbed observations (Houtekamer and Mitchell 1998). In addition, the analysis increments are perturbed by random isotropic perturbation fields (Houtekamer et al. 2009). The different sets of N_{par} parameters serve in the recentering of the analysis ensemble (Houtekamer et al. 2018) and in the configuration of the atmospheric forecast model that is used by the EnKF to advance the ensemble of analyzed states to the next analysis time.

Step 3: Comparison with observations.

At each of the N_{analysis} steps, each of the $O(1\,000\,000)$ observations that are assimilated by the EnKF (Houtekamer et al. 2018) is compared with the model trajectory of each member. The required applications of the forward operator $\mathcal{H}$ are provided by the EnKF. The CRPS is computed with these data.

Step 4: Replace parameter sets.

Using the collected results of the N_{analysis} 6-h analysis steps, replace up to N_{iter} parameter sets with the following iterative procedure:

- Compute the N_{ens} cost function values that result from removing only one member.
- Identify the badly performing member that, when removed, improves (i.e., reduces) the cost function the most.
- Identify the well performing member that, when removed, leads to the biggest degradation in cost function. It is, for now, assumed that duplicating this member and using it to replace the worst performing member will lead to an improved ensemble.
- Verify that replacement of the worst performing member with a copy of the best-performing member does indeed lead to an improved cost function value. If it does not, the iteration stops. If it does, the parameter set of the bad member is modified. For a continuous parameter p , the new value p_{new} is obtained as a perturbed and bounded

weighted sum of parameter values p_{best} and p_{worst} of, respectively, the best and worst performing members:

$$p_{\text{try}} = \alpha p_{\text{best}} + (1 - \alpha) p_{\text{worst}} + \varepsilon_p, \quad (6)$$

$$\varepsilon_p \sim N[0, (2\alpha - 2\alpha^2)\sigma_p^2], \quad (7)$$

$$p_{\text{new}} = \max[p_{\text{min}}, \min(p_{\text{max}}, p_{\text{try}})]. \quad (8)$$

Here, the variance σ_p^2 is computed from the set of N_{ens} available values of the parameter p . Note that a higher value of α would lead to a stronger contraction toward the parameter p_{best} of the best-performing member. The amplitude of the noise term [Eq. (7)] is such that, in the case of uninformed updates where the observations had no information on the parameter values, the variance σ_p^2 of the set of parameter values remains unchanged. For instance, if $p_{\text{best}} \sim N(0, \sigma_p^2)$ and $p_{\text{worst}} \sim N(0, \sigma_p^2)$ then $p_{\text{try}} \sim N(0, \sigma_p^2)$. In Eq. (8), the value is clipped to remain in the range $[p_{\text{min}}, p_{\text{max}}]$ of acceptable values. For a categorical parameter, the parameter will change to p_{best} with probability α and otherwise remain unchanged. In this study, we use the same value $\alpha = 0.9$ for updating continuous and categorical parameters.

As default value, we set $N_{\text{iter}} = 32$. Here, a parameter set is not allowed to serve multiple times in a replacement. Replacing 32 different parameter sets per day, we can potentially have substantial accumulated changes after a one-week experimental period i.e., as many as $7 \times 32 = 224$ changes. Using a value of N_{iter} that is too high would lead to over fitting of the model parameters to the weather patterns of the day and to a collapse of the parameter sets on a unique solution. With values of N_{iter} of up to 32, we observed a rather steady evolution of parameter estimates over the experimental period (e.g., Fig. 2). Higher values were not tried.

Step 5: For a new (subsequent) experiment day, continue with step 2.

A very similar algorithm is used when all members are scored individually with Eq. (5). In steps 4a–c, the worst members are simply those with the highest cost $\mathcal{J}_1$ and the best members are those with the lowest cost.

3. Modeled uncertainty

This study arose in the context of the migration of the GEPS to a new version of the global environmental multiscale (GEM) model (Côté et al. 1998a,b; Girard et al. 2014). The new model features substantial changes to the model physics component that led to major improvements in forecast quality following the summer 2019 operational upgrade of the deterministic prediction systems (McTaggart-Cowan et al. 2019b). The migration of the GEPS to the new physics was initiated later and the system was not ready for the summer 2019 upgrade. Compared to the deterministic systems, the ensemble system has additional complexity due to the multi-physics approach (Houtekamer 2011) that is used in the operational GEPS, due to the use of algorithms for stochastic physics perturbations and for stochastic kinetic energy

backscatter (Charron et al. 2010), as well as due to the use of different numbers of vertical levels in the data assimilation and medium-range forecast components.

In the migration of the GEPS system to the new model, changes are made with respect to the configuration that is considered best in the context of the 15-km-resolution Global Deterministic Prediction System (GDPS).

a. The control configuration

The future configuration of the GEPS system can benefit from an upgrade to the computational facilities of the Canadian Meteorological Centre (CMC). This permits an increase in the number of vertical levels for both the global EnKF and the medium-range forecast ensemble to 84 as in the GDPS (McTaggart-Cowan et al. 2019b), with the first prognostic thermal level at 10 m above the ground level and the first momentum level at 20 m. With respect to the operational EnKF system, this amounts to adding three vertical levels with the higher vertical resolution being in the planetary boundary layer and mid-troposphere (McTaggart-Cowan et al. 2019b, section 2.3). For the medium-range ensemble forecasts, the number of levels almost doubles, going from 45 to 84, with a much better vertical resolution both near the surface and in the stratosphere. The horizontal grid spacing of the GEPS remains at approximately 39 km as in the operational system. Following the new GDPS, a filter with a sharp $3\Delta x$ response function has been used to generate the topography field. The time step is still at 15 min.

As a starting point for our fairly expensive experiments with the evolutionary algorithm, we used the configuration of the 15-km-resolution GDPS. Some resolution-dependent parameters were changed to get a control configuration of good quality. These are: 1) the threshold vertical velocity in the trigger function of the deep convection scheme (Kain and Fritsch 1992), 2) the minimum environmental mass flux necessary for the activation of the midlevel convection scheme (McTaggart-Cowan et al. 2019b), 3) the diffusion coefficient for the diffusion layer at the top of the model, and 4) the critical gravity wave phase change in the low-level blocking scheme (H_{nc} in Lott and Miller (1997), but named “sgophic” hereinafter). The minimum midlevel convection environmental mass flux was derived as the vertical gridscale mass flux corresponding to the 95th percentile of resolved-scale midlatitude 500-hPa vertical velocities, following the same strategy used for the GDPS. The diffusion coefficient in the sponge zone was readjusted to ensure that the same portion of amplitudes of the shortest resolved waves be removed over a single time step as in the GDPS. The other two parameters were obtained by fine tuning, based on performing 10-day forecasts issued 36 h apart over two 2-month periods (July–August 2016 and January–February 2017), initialized with the GEPS ensemble mean analyses. It is worth noting that the control value of the deep convection threshold was obtained by minimizing the errors of the upper-air variables with respect to the radiosonde observations, whereas the control model performance in terms of gauge-estimated precipitation is nonoptimal. The overall sensitivity of the control model performance to the parameter sgophic was found to be rather small. Parameter values for the control configuration are given in Table 1 and discussed in more detail below.

b. Available parameters for multiphysics

In an effort to represent model error through physics parameter variations between members, parameter uncertainty corresponding to a variety of aspects of the forecast model could be sampled. However, for the initial tests with the optimization algorithm, it was thought prudent to attempt the optimization of only 10–20 parameters (Bengtsson et al. 2008). Table 1 provides our list of selected parameters with some summary information. Individual parameters are described in more detail in the appendix. As can be seen, some parameters have a continuous range of possible values whereas others are categorical. Continuous parameters can be used to fine tune a particular algorithm whereas categorical parameters permit selecting among different possible algorithms.

4. Experiments

In this section, the genetic algorithm discussed in section 2 is tested with the multiphysics described in section 3. In setting up for the experiments, a humidity bias in the EnKF was identified and dealt with as described in section 4a. In section 4b, the experiments are described. In section 4c, the parameter ensembles are allowed to evolve with the genetic algorithm. The impact on the quality of short-range forecasts is evaluated in section 4d. Correlations between parameter estimates are the subject of section 4e and the impact on medium-range ensemble forecasts is discussed in section 4f. Finally, section 4g provides verification against global precipitation data.

a. Removal of a humidity bias in the EnKF

The experiments documented in this study start from a research version of the operational Canadian global EnKF system that uses the updated GEM model described in section 3. Initial experiments with the EnKF were encouraging, but showed an unexpected strong correlation between the parameters for the analysis recentering and for deep convection. From the point of view of forecast model development, the correlation of model parameters with data assimilation parameters is not desirable. One would like to develop a model that behaves like the atmosphere, independent of any shortcomings of the data assimilation systems. Furthermore, if information from the data assimilation cycle is to be used to adjust the forecast model, as we propose in this study, the data assimilation process itself should not be a source of bias.

In Houtekamer et al. (2018, section 5.1), it was noted that the removal of supersaturation, when adding isotropic perturbations to analyzed states, could act as a source of bias. Selectively removing humidity, when the analyzed ensemble is in nearly saturated conditions, eventually leads to a cold temperature bias. In our current operational and research configurations of the EnKF, the Incremental Analysis Update (IAU) procedure (Bloom et al. 1996) is used to add analysis increments in a gradual manner to the background estimate. Here, supersaturated analyzed states are not necessarily problematic because the corresponding increments are gradually added to the model trajectory. In the model, precipitation serves as the natural process to limit the amount of humidity in the atmosphere. It was thus decided to remove the test for supersaturation

TABLE 1. The parameters that are optimized in this study. The parameters are explained in some detail in the [appendix](#). The value found using prior deterministic experiments is given in the “control” column ([section 3a](#)). For a parameter value, one has either a small set of different options or a continuous range of values as indicated by the last two columns.

Identifier	Short description	Control	Distribution	Permitted range/values
γ	Weight for EnVar in the analysis recentering	Does not apply	Continuous	[0, 1]
kfc2/kfc3	Choice of Kain–Fritsch (KF) implementation	kfc2	Categorical	kfc2, kfc3
kfcrad	Updraft radius for KF over land	1500 m	Continuous	[1300, 1700] m
kferadw	Updraft radius for KF over water	1250 m	Continuous	[800, 1300] m
kfc_trigger	Trigger parameter for KF	0.05 m s^{-1}	Continuous	[0.03, 0.08] m s^{-1}
bkf_closures	Closure for shallow convection	equilibrium	Categorical	equilibrium, CAPE
bkf_evaps	Evaporation of detrained condensate	false	Categorical	false, true
mixing	Mixing length estimate	turboujo	Categorical	black62, turboujo, boujo
sl_lmin_soil	Minimum Obukhov length over land	20 m	Continuous	[5, 20] m
fnn_reduc	Reduction of flux enhancement from PBL clouds	0.8	Continuous	[0.5, 1.0]
func_stab	Surface layer stability function	beljaars91	Categorical	beljaars91, delage97
sgophic	Nondimensional subgrid-scale mountain height	0.15	Continuous	[0.0, 0.5]
rad_cond_rei	Effective radius of ice particles	$15 \mu\text{m}$	Continuous	[15, 35] μm

after the addition of the isotropic perturbations. With this change, the correlation between the parameter for the scheme for deep convection and the parameter for the analysis recentering became small with values in the range $[-0.19, 0.27]$ for the four tuning experiments of the manuscript ([Table 2](#)).

b. Description of the experiments

All experiments were started at 1800 UTC 27 December 2016, with a 256-member ensemble of short-range forecasts. The first EnKF analysis is valid 0000 UTC 28 December 2016. Changes to the parameter settings are made every 24 h, i.e., using $N_{\text{analysis}} = 4$, starting after the analysis of 0000 UTC 29 December 2016, and continuing for 23 days until 0000 UTC 20 January 2017.

The experiments are listed in [Table 2](#). Whereas the current manuscript discusses an algorithm to estimate parameters, the study was performed in the context of a system upgrade at the CMC. To gauge the relative importance of the parameter changes, as compared to a major system change at CMC, we include verifications from the so-called “Tuned CMC hybrid” that had been obtained in the context of the development of the hybrid gain algorithm ([Houtekamer et al. 2018](#)) and is further described in that manuscript.

The experiment STABLE serves as a starting point for the optimization experiments. Here, parameter ensembles are used in a multiphysics approach, but these ensembles are kept unchanged over the experimental period. This experiment can thus serve as a baseline for experiments that do update the parameter ensembles. The migration to new model physical parameterizations is a major component of the difference between the Tuned CMC hybrid and the STABLE experiments.

The experiment COST uses the score for individual members to optimize performance [Eq. (5)], whereas experiment CRPS uses the ensemble score [Eq. (4)].

To gauge the dependence of the results on the initial parameter sets, experiments COST-PER and CRPS-PER start from initial parameter ensembles whose values were specified slightly differently. Experiments COST and CRPS obtain the initial ensemble for continuous parameters using a normal distribution centered on the middle of the acceptable interval,

where values beyond the limits are clipped. Experiments COST-PER and CRPS-PER, use the square of a random uniform value between 0 and 1 to obtain the scaled distance of the parameter value of an ensemble member from the minimum of the range of parameter values. Taking the square will shift the values toward the minimum value. The impact on the mean and standard deviation can be seen in [Table 3](#). (As a further illustration, for the parameters kfc_trigger and fnn_reduc, the clouds of initial parameter values are shown in [Fig. 6a](#).) For categorical parameters, experiment COST and CRPS assign equal probability to all possible values. Experiments COST-PER and CRPS-PER will assign higher probability to the first listed values (in [Table 1](#)) of the categorical parameters. The impact is again visible in [Table 3](#).

As in [Houtekamer et al. \(2018\)](#), 5 day forecasts with the GEPS are used to evaluate the quality of experimental configurations. However, the GEPS also benefits from the optimized parameter sets as they become available from the genetic algorithm. This addresses concerns related to inconsistencies of the data assimilation and forecast phases ([Ollinaho et al. 2013](#); [Trenberth and Guillemot 1998](#)). It is to be noted that the “Tuned CMC hybrid” uses only 45

TABLE 2. List of experiments.

Name	Short description
Tuned CMC-hybrid	The best-performing configuration with the previous generation of model physics (Houtekamer et al. 2018)
STABLE	Multiphysics are used and kept unchanged during the experimental period
COST	Multiphysics are used and allowed to evolve based on a score for individual members
COST-PER	As in COST, but with initially perturbed parameter values
CRPS	Multiphysics are used and allowed to evolve based on an ensemble score
CRPS-PER	As in CRPS, but with initially perturbed parameter values

TABLE 3. Parameter values at the beginning of the experiments. Experiments COST and CRPS start with the same default samples of parameter values. Experiments COST-PER and CRPS-PER start from the perturbed set of values.

Identifier	Default	Perturbed
γ	0.52 ± 0.26	0.35 ± 0.31
kfc2/kfc3	{50%, 50%}	{65%, 35%}
kferad	1502 ± 108 m	1439 ± 127 m
kferadw	1053 ± 126 m	989 ± 158 m
kfc_trigger	0.056 ± 0.013 m s ⁻¹	0.047 ± 0.015 m s ⁻¹
bkf_closures	{49%, 51%}	{70%, 30%}
bkf_evaps	{46%, 54%}	{66%, 34%}
mixing	{28%, 35%, 37%}	{52%, 37%, 11%}
sl_lmin_soil	12.4 ± 4.4 m	9.7 ± 4.5 m
fnn_reduc	0.759 ± 0.104	0.649 ± 0.142
func_stab	{47%, 53%}	{66%, 34%}
sgophic	0.258 ± 0.098	0.176 ± 0.149
rad_cond_rei	24.9 ± 5.4 μ m	21.6 ± 5.9 μ m

vertical levels for the medium-range forecasts, whereas the other experiments benefit from using 84 vertical levels.

c. Evolution of parameter estimates

The results of the tuning experiments are summarized in Table 4. For the γ parameter, the final mean value is near 0.5 for the four experiments. However, the final std dev is much smaller for the pair of COST experiments than for the pair of CRPS experiments. Histograms for the distribution at the final time show bimodality for the CRPS experiments, with many values near the limit of the [0–1] range of the parameter (Fig. 1). Such distributions were previously recommended by Houtekamer et al. (2018) with the name “CMC-hybrid.” Apparently, the CRPS score is able to give value to sampling analysis differences with the γ parameter. Generally, we note smaller standard deviations for the COST experiments than for the corresponding CRPS experiments. The only two exceptions are for the default experiments, where we note a somewhat larger spread for the sgophic and rad_cond_rei parameters in the COST experiment. This illustrates that the deterministic score permits zooming in on the value that gives

best results on average whereas the ensemble score attempts to maintain a range of values that covers the possible results reliably.

Intriguing results were obtained for the mixing length estimate (Fig. 2). For the four experiments, the “black62” parameterization ends up being used by the majority of members. In the COST experiments, the “boujo” scheme is almost completely eliminated. Some follow-up experiments with longer-range forecasts confirmed the good performance of the black62 scheme. These results are different from those obtained with the GDPS system at our center (McTaggart-Cowan et al. 2019b), which may either be due to a resolution dependence or to using different verification procedures.

The evolution of the estimates for the kfc_trigger parameter is shown in Fig. 3. Overall, the changes in the probability distribution are rather small. However, both for the COST and CRPS experiments, it would seem that the distance between the distributions of the default and the initially perturbed ensembles reduces with time. The possible convergence to a common value suggests that the model trajectories are somewhat sensitive to this parameter. For the COST experiments, there seems to be convergence, with reducing standard deviations, toward a value that is lower than the initial estimate that was centered on 0.056 m s⁻¹. The CRPS experiments show no evidence of convergence to a different value.

Troubling behavior is seen for the estimates of the fnn_reduc parameter in Fig. 4. Here the distance between the default and the perturbed ensembles does not reduce with time (cf. Tables 3 and 4). At the end of the 20-day period, the box-whisker inner bars are no longer overlapping. The observed reductions in std dev would seem to be not appropriate in this case. This is one example of a parameter for which the currently proposed algorithm does not manage to extract useful information from the observations.

d. Verification of short-range forecasts

At our center, the comparison of model forecasts with radiosonde observations is used in the evaluation of potential changes to the NWP system. Such a comparison is, for instance, shown in Houtekamer et al. (2018, Table 4), where various changes to the system lead to desirable reductions of

TABLE 4. Estimated parameter values at the end of the experiments.

Identifier	COST	COST-PER	CRPS	CRPS-PER
γ	0.48 ± 0.08	0.46 ± 0.09	0.55 ± 0.35	0.53 ± 0.36
kfc2/kfc3	{44%, 56%}	{75%, 25%}	{41%, 59%}	{36%, 64%}
kferad	1480 ± 67 m	1480 ± 98 m	1523 ± 95 m	1511 ± 111 m
kferadw	931 ± 80 m	887 ± 67 m	1010 ± 136 m	988 ± 136 m
kfc_trigger	0.047 ± 0.009 m s ⁻¹	0.042 ± 0.008 m s ⁻¹	0.058 ± 0.012 m s ⁻¹	0.052 ± 0.012 m s ⁻¹
bkf_closures	{68%, 32%}	{96%, 4%}	{59%, 41%}	{84%, 16%}
bkf_evaps	{29%, 71%}	{79%, 21%}	{64%, 36%}	{75%, 25%}
mixing	{79%, 18%, 4%}	{62%, 38%, 0%}	{68%, 14%, 18%}	{68%, 16%, 16%}
sl_lmin_soil	13.5 ± 3.8 m	10.2 ± 3.2 m	13.7 ± 4.0 m	10.5 ± 3.6 m
fnn_reduc	0.796 ± 0.065	0.665 ± 0.084	0.814 ± 0.088	0.680 ± 0.118
func_stab	{57%, 43%}	{45%, 55%}	{64%, 36%}	{65%, 35%}
sgophic	0.192 ± 0.101	0.183 ± 0.093	0.234 ± 0.092	0.198 ± 0.152
rad_cond_rei	25.4 ± 4.3 μ m	22.9 ± 4.3 μ m	23.1 ± 4.2 μ m	21.1 ± 4.8 μ m

FIG. 1. Histograms for the γ parameter at the final time for the COST, COST-PER, CRPS, and CRPS-PER experiments as identified in the legend.

innovation amplitudes of 1.4%–1.9%. In Table 5, we similarly compare the Tuned CMC-Hybrid, COST, COST-PER, CRPS, and CRPS-PER experiments with the STABLE experiment. The verifying data are for the period from 0000 UTC 1 January 2017 to 1200 UTC 20 January 2017. A very small degradation of -0.073% is noted for the “COST-PER” experiment, mainly due to a degradation in the dewpoint depression variable,

whereas there are small improvements of 0.035%, 0.048% and 0.142% for, respectively, the COST, CRPS, and CRPS-PER experiments. These differences are approximately an order of magnitude smaller than the differences between the Tuned CMC-hybrid and STABLE algorithm and thus do not permit conclusions on the relative quality of the different optimization methods.

Small systematic changes in quality were found for radiance channels that are sensitive to humidity, where the optimization experiments show a slightly reduced bias. This is shown for AMSU-B observations in Fig. 5. Similar verifications were obtained for humidity sensitive ATMS and infrared channels.

e. Correlation between parameter estimates

The impact of different parameters need not be orthogonal. There are, for instance, four parameters for deep convection in the Kain–Fritsch parameterization. It is to be expected that the optimal values of these four parameters are correlated. The genetic algorithm should be able to handle this situation as it can remove parameter sets that lead to bad performance and duplicate parameter sets that contribute to good ensemble performance.

At the end of each experiment, there are 256 sets of N_{par} parameters. The correlations were computed for the $(1/2)[N_{\text{par}} \times (N_{\text{par}} - 1)]$ possible pairs of different parameters. In the case of categorical parameters, parameter values were represented

FIG. 2. Results of the experiments COST, COST-PER, CRPS, and CRPS-PER for the mixing length algorithm. For each of the three possible algorithms (black62, turboujo, and boujo), the percentage of members that use the algorithm is shown as a function of the experiment time. Percentages are shown every other day.

FIG. 3. Box-and-whisker plot for the *kfc_trigger* parameter. The boxes extend from the first to the third quartile. The median value is shown with a horizontal line. Points outside the box by more than 1.5 times the interquartile distance are marked as outliers. Results are shown every other day.

by 0 and 1 in the case of two possible values and by 0,1 and 2 in the case, for the mixing length estimate, with 3 possible values. For experiment COST, COST-PER, CRPS, and CRPS-PER, correlations were, respectively, in the ranges $[-0.35, 0.43]$, $[-0.34, 0.36]$, $[-0.33, 0.50]$, and $[-0.4, 0.37]$. Correlations between the estimates of different parameters were thus small and in absolute value below 0.5. For the parameters *kfc_trigger* and *fnn_reduc*, the clouds of

parameter combinations are shown in Fig. 6b. For the 4 experiments, the correlations between these parameters are 0.03, -0.03 , 0.14, and -0.04 . It is noted that the pairs of values continue to cover a large fraction of the available space, as is desirable for continued functioning of the optimization algorithm, as opposed to clustering on a small number of different values. In view of the current good coverage, it could be considered to increase the contraction

FIG. 4. As in Fig. 3, but for the *fnn_reduc* parameter.

TABLE 5. Percentage improvement in 6-h verifications of ensemble-mean states are given with respect to the STABLE experiment. The verification is against independent observations from approximately 600 radiosonde stations for the period 0000 UTC 1 Jan–1200 UTC 20 Jan 2017. Positive values indicate a higher-quality system. Results are given for zonal wind u , meridional wind v , (geopotential) height, temperature T , and dewpoint depression $T - T_d$. For wind, height, and temperature, the values are averages over verifications at 10, 20, 30, 50, 70, 100, 150, 200, 250, 300, 400, 500, 700, 850, 925, and 1000 hPa. For dewpoint depression, observations are at 250, 300, 400, 500, 700, 850, 925, and 1000 hPa.

Variable	Tuned CMC-hybrid	COST	COST-PER	CRPS	CRPS-PER
u	−1.190	0.068	0.069	0.046	0.125
v	−1.399	0.075	0.153	−0.004	0.167
Height	−0.128	0.122	−0.049	0.126	0.180
T	−0.203	0.049	0.117	0.030	0.280
$T - T_d$	−0.545	−0.139	−0.652	0.046	−0.041
Average	−0.693	0.035	−0.073	0.048	0.142

parameter α in Eq. (6), as this would permit a reduction of the amplitude of the noise term.

In practice, when preparing an NWP system for operational use, it is important to be aware of possible compensating impacts of different parameters. It would appear that the currently used parameters have mostly independent impacts. This is a desirable situation, in which parameters can be studied in isolation. It may, however, be a consequence of the particular form of the update equation [Eq. (6)] in which noise is added independently for different parameters. A reduction of the noise term, as suggested above, would favor the detection of

possible correlations and bimodality. An alternative, aiming only at the detection of correlations, would be to sample the noise from an evolving multidimensional Gaussian distribution as in the original EPPES system.

f. Impact on medium-range forecasts

For 20 days, from 0000 UTC 1 January to 1200 UTC 20 January 2017, 5-day ensemble forecasts are performed twice a day with the evolved parameter ensembles that are available at that time. The impact of the parameter changes on 5-day ensemble forecast quality is summarized in Table 6.

FIG. 5. Improvement in bias as observed with AMSU-B observations.

FIG. 6. Scatterplot for the $kfc_trigger$ and fnn_reduc parameters at the (a) initial and (b) final times of the experiment. Note that experiments COST and COST-PER are not shown in (a) because values are identical to those of experiments CRPS and CRPS-PER, respectively.

For all experiments, we note a substantial reduction of std dev of more than 7% when compared with the “Tuned CMC-hybrid.” This can be mostly attributed to the improved model physics and the increased number of vertical levels. We note an additional modest improvement of about 0.3% for experiments COST, COST-PER, and CRPS and a larger improvement of order 0.9% for experiment CRPS-PER. The improvement in average score is mostly for the geopotential height variable. Further inspection showed a reduction of biases for geopotential height. Such changes, albeit small, were hoped for given the changed humidity biases in the assimilation cycle (Fig. 5).

In Table 7, the improvement is given for a few latitude bands. From the STABLE experiment, we note that the impact of the new physical parameterizations is biggest in the extratropical regions. Whereas the optimization experiments all provide some further improvement, this is most noteworthy for the southern extratropics. We speculate that, our NWP center itself being in the Northern Hemisphere, the potential for model improvement is biggest in the southern extratropics.

In Table 8, the improvement is given as a function of forecast time. From the STABLE experiment, we note the growing impact of the model improvement with forecast time, with a saturation of the improvement near day 3. Such a cumulative impact is likely proper to model improvements. Indeed, we observe a similar growing impact when comparing the optimization experiments with the STABLE experiments. Note, for instance, that the improvement at day 5 is bigger than the improvement at day 1 for all experiments.

g. Verification of precipitation against the GPM IMERG product

The precipitation rates forecast by the different experiments were verified against the IMERG product made available by

NASA’s Precipitation Processing Center (Huffman et al. 2014). This product provides gauge-calibrated estimates of precipitation rates by combining information from multiple satellites. It is available on a global grid with a resolution of $0.1^\circ \times 0.1^\circ$ every 30 min. In the context of this study, it is interesting to use IMERG since it provides information that is independent from all the observations that are assimilated.

Prior to the verification, the IMERG precipitation rates were interpolated to the “Yin” and “Yang” grids of the model by averaging all observations found within a radius of 40 km from the center of each grid tile. This spatial smoothing was primarily applied to remove the small-scale features that are present in the observations but that the model is not capable of representing. The averaging process also makes the observed distribution of precipitation closer to the distribution simulated by the model.

As an example, the modeled and IMERG precipitation rates are shown in Figs. 7a and 7b, respectively. The gridded

TABLE 6. Percentage improvement in 5-day GEPS CRPS verifications with respect to the tuned CMC-hybrid. For the u and v wind components, temperature T , and geopotential height, individually printed values have been obtained as the average improvement at the levels 10, 50, 100, 250, 500, 850, and 925 hPa. For dewpoint depression, the average is only over the levels 250, 500, 850, and 925 hPa. The verification is against radiosonde observations.

	STABLE	COST	COST-PER	CRPS	CRPS-PER
u	10.474	10.534	10.654	10.404	10.867
v	9.477	9.736	9.620	9.320	9.881
Height	6.719	7.658	7.414	7.968	9.248
T	8.663	9.016	9.135	8.711	9.515
$T - T_d$	1.738	2.034	2.022	2.013	2.287
Average	7.414	7.796	7.769	7.683	8.360

TABLE 7. As in table 6, but the results have, in addition, been averaged over the five verified variables and are given for different geographical regions: global, northern extratropical (latitude $> 20^\circ$), tropical ($-20^\circ < \text{latitude} \leq 20^\circ$), and southern extratropical (latitude $\leq -20^\circ$).

	STABLE	COST	COST-PER	CRPS	CRPS-PER
Global	7.414	7.796	7.769	7.683	8.360
Northern extratropics	7.567	7.901	7.831	7.585	8.246
Tropics	3.504	3.714	3.615	4.155	4.121
Southern extratropics	7.157	8.126	8.414	9.043	10.344

precipitation estimates of the IMERG product (PR) are accompanied by a quality index (QI) ranging between zero and one, with one indicating observations of the highest quality. During the interpolation process, the N_i observations of PR and QI found in the vicinity of a given model grid tile were averaged using

$$\text{PR}_{\text{avg}} = \frac{\sum_{i=1}^{N_i} \text{QI}_i \text{PR}_i}{\sum_{i=1}^{N_i} \text{QI}_i}, \quad (9)$$

$$\text{QI}_{\text{avg}} = \frac{\sum_{i=1}^{N_i} \text{QI}_i \text{QI}_i}{\sum_{i=1}^{N_i} \text{QI}_i}, \quad (10)$$

respectively. The weighting by QI ensures that the precipitation estimates with the highest-quality contribute most to the resulting average. An example of the interpolated quality index is shown in Fig. 7c.

Six-hour trials from all experiments were verified four times per day for a period extending between 1 and 20 January 2017. The Continuous Rank Probability Score (CRPS) was then computed following (Hersbach 2000) and used as the verification metric. To mitigate the very skewed distribution of precipitation rates, the CRPS was estimated using the log-transformed quantity $\log(\text{precip_rate} + 0.1)$. For precipitation simulated by the global model, this quantity ranges approximately between -2.3 and 3.5 (0 and $\sim 33 \text{ mm h}^{-1}$, respectively). For the computation of the CRPS, observations were weighted by their quality index (as per section 4c of Hersbach 2000) in order for the better observations to have a greater influence on the verification results.

Figure 8 shows the average difference between the CRPS of experiments COST, CRPS, COST-PER, and CRPS-PER and the CRPS of the reference experiment STABLE. For concise results, the CRPS differences obtained during the entire verification period were aggregated together. A bootstrap approach with 1000 samples was used to estimate the 90% confidence intervals for the average difference in CRPS.

In terms of total CRPS, both the CRPS and CRPS-PER experiments improve precipitation forecasts in the Northern and Southern hemispheres. These improvements are mostly caused by improvement in the reliability component of the CRPS score. The COST and COST-PER experiments do not perform as well and generally degrade precipitation forecasts.

Tropical precipitation is much more difficult to predict. All experiments, except CRPS with neutral results, deteriorate precipitation forecasts in this region.

5. Discussion and conclusions

In this study, we have explored the use of a genetic algorithm to estimate model parameters in a data assimilation context. For many of the parameters, the pairs of experiments, which started from different initial parameter sets, did not converge to similar distribution functions over the 24-day experimental period. This suggests that the observational network, as currently used, does not provide enough specific information on these parameters to improve upon the initial distributions provided by the experts on model physical parameterizations. In future experiments with the genetic algorithm, we intend to use specially generated ensembles of 24-h forecasts started from the initial conditions provided by the EnKF. The rationale is that the requirement to shadow observations for 24 h is much more constraining on the quality of the model than shadowing for only 6 h (Orrell et al. 2001). Another possibility would be to include either all-sky microwave radiances (Zhu et al. 2016) or the above used GPM IMERG product in the evaluation function to arrive at a more representative sampling of possible atmospheric conditions with the verifying observations.

In the case of the parameter γ for the weight in the hybrid gain algorithm, the estimated probability distribution was close to a previously proposed bimodal distribution of the CMC-hybrid (Houtekamer et al. 2018), for the two experiments CRPS and CRPS-PER that used an ensemble score to estimate the value of an individual parameter set (see Fig. 1). This contrasts with experiments COST and COST-PER that scored individual members and arrived at a narrow unimodal distribution for γ .

TABLE 8. As in table 6, but the results have, in addition, been averaged over the five verified variables and are given per forecast length: day 1, day 2, ..., day 5.

	STABLE	COST	COST-PER	CRPS	CRPS-PER
Days 1–5	7.414	7.796	7.769	7.683	8.360
Day 1	6.204	6.436	6.477	6.313	6.792
Day 2	6.815	7.152	7.211	7.115	7.733
Day 3	7.722	8.079	8.052	8.084	8.782
Day 4	8.235	8.766	8.592	8.569	9.358
Day 5	8.095	8.545	8.513	8.335	9.134

FIG. 7. Visual comparison of the precipitation rates (a) simulated in the experiment STABLE, and (b) from IMERG interpolated onto the model grid. (c) The quality index associated with the IMERG product at different locations. All quantities are valid at 0000 UTC 3 Jan 2017.

For the algorithm to estimate mixing length, the four experiments favor the Blackadar (1962) formulation, referred to as “black62.” Some recent sensitivity tests run independently with the GDPS suggest that the use of black62 may indeed improve forecast skill over the oceans. The results of the genetic algorithm may be consistent with those sensitivity tests.

The genetic algorithm can identify correlations between parameters and thus compensate errors by the simultaneous adjustment of multiple parameters. Such a correlation was found during preparatory experiments and related to a humidity bias in the EnKF system that could be removed. In the current experimental configuration, correlations between parameters were small. This may be partly related to the

addition of fairly large amplitude noise, that is independent for different model parameters, when updating the parameter values [Eq. (6)]. Note, however, that the absence of correlations may also be related to the limited sensitivity of the verification system to changes in the parameter values.

A puzzling result of our study is the small amplitude of the improvements in scores that were obtained with the genetic algorithm. The error reduction of 0.1% that is directly due to the genetic algorithm is an order of magnitude smaller than had been seen with other operational upgrades to the EnKF system at our center. Perhaps major upgrades are more naturally associated with increased resolution or with an improved or new insight of the behavior of the NWP system as opposed to diligent optimization within an already existing framework. We also note that the optimization started from a physics package that had already been optimized using different procedures such that the remaining room for improvement was perhaps small. Another possibility is that physical parameterizations are too simplistic to accurately represent the real complexity of atmospheric processes. Consequently, varying the adjustable parameters of the parameterizations cannot represent the real variability of the actual atmospheric processes that we would like to model. Whereas the evolutionary algorithm can deal with structural uncertainty by means of categorical parameters, it would require a substantial effort to develop a relevant hypothesis space.

For most parameters, the use of an ensemble score resulted in larger standard deviations for the parameter ensembles than corresponding experiments with a deterministic score. It thus appears that the ensemble score can protect against underdispersion of the parameter ensembles. An independent verification against the GPM IMERG product did indeed show improved reliability for the experiments using the ensemble score.

Optimization of parameters in the data assimilation context, coincided with small improvements in both the data assimilation and medium-range forecasts contexts. Such behavior is consistent with a seamless forecasting system (Bauer et al. 2015) and is highly desirable. However, there is no guarantee that parameters that perform well in the data assimilation context of the EnKF will also perform well in the context of the medium-range forecasts of the GEPS. In fact, Schirber et al. (2013) find that estimated parameter values that lead to an error reduction on short time scales can lead to an increase of error on climatological time scales. Our feeling is that the data assimilation environment can be used to optimize the success in predicting the short time scales of the daily weather whereas longer-range forecasts are required to also guarantee an appropriate behavior for the longer time scales that determine the model climate.

In the future, based on our experience in the context of a major upgrade to the model’s physical parameterizations, we expect to use the evolutionary algorithm to support the initial phases of research projects. We are currently considering the use of a new land surface system with the GEPS, and the evolutionary algorithm could serve to explore the response of both the surface and the atmospheric system to the coupling. Such work would be done in close collaboration with experts in

FIG. 8. Verifications results of log-transformed precipitation rates estimated against the GPM IMERG product. This chart represents the change in total CRPS associated with the different parameter adjustment strategies (COST, CRPS, COST-PER, and CRPS-PER) relative to the STABLE experiment. Average results for the total CRPS and its decomposition into reliability and potential are presented for the Northern Hemisphere (blue), the tropics (orange), and the Southern Hemisphere (green). For all panels, negative values indicate improvement with respect to the STABLE experiment. Color shadings indicate the 90% confidence interval for the mean difference in CRPS (solid line).

both systems. We do not expect that the evolutionary algorithm will be used in an autonomous manner in the operational ensembles. It would be hard to prove that the model physics cannot drift to unphysical suboptimal settings, and settings optimized during one season might be inappropriate for subsequent seasons. Our current experience also suggests that experts, given the guidance provided by the algorithm, often arrive at more appropriate changes in related components of the system.

Acknowledgments. A discussion with Sebastian Reich on the role of using an ensemble score helped to clarify the experimental design. The IMERG data were provided by NASA's Precipitation Processing Center and PPS, which develop and compute the GPM IMERG product as a contribution to GPM, and archived at the NASA GES DISC. We thank Herschel Mitchell and two anonymous reviewers for their thoughtful reviews of the manuscript.

APPENDIX

Parameters for Optimization

a. Analysis recentering

Starting with the summer 2019 operational upgrade, the CMC is using an operational hybrid gain algorithm in the

global EnKF (Penny 2014; Bonavita et al. 2015; Houtekamer et al. 2018). Here the EnKF analyses are recentered on the mean of the EnKF and the Ensemble Variational (EnVar) analyses to give an ensemble of hybrid analyses $x_{i,\text{hyb}}^a$, $i = 1, \dots, 256$. The correction of the original ensemble of EnKF analyses $x_{i,\text{EnKF}}^a$, $i = 1, \dots, 256$ toward the EnVar analysis x_{EnVar}^a is controlled by the parameter γ :

$$x_{i,\text{hyb}}^a = x_{i,\text{EnKF}}^a + \gamma_i (x_{\text{EnVar}}^a - \overline{x_{i,\text{EnKF}}^a}), \quad i = 1, \dots, 256. \quad (\text{A1})$$

At the beginning of the iterative optimization, the 256 members have different correction factors γ_i with a minimum of 0 (no correction toward the EnVar) and a maximum of 1 (fully recentering on the EnVar). The optimization algorithm is meant to provide the pdf of the optimal distribution for γ . In the operational ensemble, all members use $\gamma_i = 0.5$.

b. Deep convection scheme

The Kain and Fritsch parameterization (Kain and Fritsch 1990, 1992) is used to represent the effects of deep convection. Some local modifications to the parameterization are described in McTaggart-Cowan et al. (2019a). In the default implementation (named “kfc2” in this manuscript), the mass flux equations are solved semi-implicitly. In the alternative “kfc3” implementation, the solution is implicit, which leads to improved convergence of the solution.

Initially, ensemble members will randomly be assigned either the kfc2 or the kfc3 implementation with equal probability.

c. Deep convection radius

The parameters “kfcrad” and “kfcradw” determine the initial radius of the convective updraft in the deep convection scheme over land and ocean, respectively. The entrainment of environmental air into the convective updraft (dilution) is inversely proportional to this value and strongly influences how deep the convective cloud will be. The range for land-based convection is 1300 to 1700 m, while values for the ocean vary from 800 to 1300 m (McTaggart-Cowan et al. 2019b, section 3.6.1).

d. Deep convection initiation

The first task of a deep convection scheme is to determine whether deep convection should be expected in the column. In the Kain and Fritsch (1990) scheme, a thermal perturbation represents the effects of subgrid-scale thermals on parcel buoyancy. This perturbation is a function of the difference between the gridscale vertical velocity and a reference vertical velocity known as the trigger parameter. Over land it is a constant, while over water it is function of the convective velocity scale as described in McTaggart-Cowan et al. (2019a). As a general rule, lower threshold values favor the triggering of convection. The range for this parameter is $[0.03, 0.08] \text{ ms}^{-1}$.

e. Closure for shallow convection

Two closures are available for the shallow convection scheme. The iterative CAPE (convective available potential energy)-based closure as described in Bechtold et al. (2001) or the direct closure based on an assumption of quasi-equilibrium (McTaggart-Cowan et al. 2019b, section 3.6.2). The two options are respectively referred to as “cape” and “equilibrium.”

f. Evaporation of detrained condensate

This switch determines whether the condensate detrained by shallow convection clouds is passed to the microphysics scheme (becoming gridscale clouds) or treated internally by the scheme itself (evaporated). The treatment by the gridscale scheme is activated by setting a switch named “bkf_evaps” to false, while internal evaporation is computed when this switch is set to true.

g. Mixing length estimate

The mixing length can be computed using the formulation of Blackadar (1962), or of Bougeault and Lacarrère (1989) or, in a combination of both, using the latter scheme in turbulent regimes and the former otherwise (McTaggart-Cowan et al. 2019b, section 3.5). These three schemes are referred to respectively as “black62,” “boujo,” and “turboujo.”

h. Minimum Obukhov length over land

A minimum value is imposed on the Obukhov length to avoid an effective decoupling of the soil surface and the atmosphere

under very stable and light wind conditions (McTaggart-Cowan et al. 2019b, section 3.4.1). Here a range of $[5, 20] \text{ m}$ is given to this parameter, which is named “sl_lmin_soil.”

i. Reduction of flux-enhancement factor from PBL clouds

A reduction factor, named “fnn_reduc,” is applied to the parameter $f_N N$ for the turbulent flux enhancement due to boundary layer clouds (Fig. 4c in Bechtold and Siebesma 1998). Appropriate values of the factor are thought to be in the range $[0.5, 1.0]$.

j. Surface layer stability functions

The stability functions due to either Beljaars and Holtslag (1991) or Delage (1997) can be used for the surface layer in the stable case (McTaggart-Cowan et al. 2019b, section 3.4.1). These scheme are named “beljaars91” and “delage97,” whereas the model option is referred to as “func_stab.”

k. Nondimensional subgrid-scale mountain height

A parameter “sgophic” appears in the parameterization for the subgrid-scale orographic drag, (H_{nc} in Eq. (9), Lott and Miller 1997). The parameter is dimensionless and may be interpreted as a scaled height that controls the depth of the blocked layer. It is initialized with the range of values $[0, 0.5]$, which is thought to be physically reasonable.

l. Effective radius of ice particles

In the radiation scheme, the parameter “rad_cond_rei” is the effective radius of ice particles (McTaggart-Cowan et al. 2019b, section 3.8). This quantity has a large impact on radiative absorption and is thought to range from 15 to 35 μm .

REFERENCES

- Aksoy, A., F. Zhang, and J. W. Nielsen-Gammon, 2006: Ensemble-based simultaneous state and parameter estimation in a two-dimensional sea-breeze model. *Mon. Wea. Rev.*, **134**, 2951–2970, <https://doi.org/10.1175/MWR3224.1>.
- Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, <https://doi.org/10.1038/nature14956>.
- Bechtold, P., and P. Siebesma, 1998: Organization and representation of boundary layer clouds. *J. Atmos. Sci.*, **55**, 888–895, [https://doi.org/10.1175/1520-0469\(1998\)055<0888:OAROB>2.0.CO;2](https://doi.org/10.1175/1520-0469(1998)055<0888:OAROB>2.0.CO;2).
- , E. Bazile, F. Guichard, P. Mascart, and E. Richard, 2001: A mass-flux convection scheme for regional and global models. *Quart. J. Roy. Meteor. Soc.*, **127**, 869–886, <https://doi.org/10.1002/qj.49712757309>.
- Beljaars, A. C., and A. A. Holtslag, 1991: Flux parameterization over land surfaces for atmospheric models. *J. Appl. Meteor.*, **30**, 327–341, [https://doi.org/10.1175/1520-0450\(1991\)030<0327:FPOLSF>2.0.CO;2](https://doi.org/10.1175/1520-0450(1991)030<0327:FPOLSF>2.0.CO;2).
- Bengtsson, T., P. Bickel, and B. Li, 2008: Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. *Inst. Math. Stat. Collect.*, **2**, 316–334, <https://doi.org/10.1214/193940307000000518>.
- Blackadar, A. K., 1962: The vertical distribution of wind and turbulent exchange in a neutral atmosphere. *J. Geophys. Res.*, **67**, 3095–3102, <https://doi.org/10.1029/JZ067i008p03095>.

- Bloom, S. C., L. L. Takacs, A. M. da Silva, and D. Ledvina, 1996: Data assimilation using incremental analysis updates. *Mon. Wea. Rev.*, **124**, 1256–1271, [https://doi.org/10.1175/1520-0493\(1996\)124<1256:DAUIAU>2.0.CO;2](https://doi.org/10.1175/1520-0493(1996)124<1256:DAUIAU>2.0.CO;2).
- Bonavita, M., M. Hamrud, and L. Isaksen, 2015: EnKF and hybrid gain ensemble data assimilation. Part II: EnKF and hybrid gain results. *Mon. Wea. Rev.*, **143**, 4865–4882, <https://doi.org/10.1175/MWR-D-15-0071.1>.
- Bougeault, P., and P. Lacarrère, 1989: Parameterization of orography-induced turbulence in a mesobeta-scale model. *Mon. Wea. Rev.*, **117**, 1872–1890, [https://doi.org/10.1175/1520-0493\(1989\)117<1872:POOITI>2.0.CO;2](https://doi.org/10.1175/1520-0493(1989)117<1872:POOITI>2.0.CO;2).
- Charron, M., G. Pellerin, L. Spacek, P. L. Houtekamer, N. Gagnon, H. L. Mitchell, and L. Michelin, 2010: Toward random sampling of model error in the Canadian Ensemble Prediction System. *Mon. Wea. Rev.*, **138**, 1877–1901, <https://doi.org/10.1175/2009MWR3187.1>.
- Côté, J., J.-G. Desmarais, S. Gravel, A. Méthot, A. Patoine, M. Roch, and A. Staniforth, 1998a: The operational CMC-MRB Global Environmental Multiscale (GEM) model. Part II: Results. *Mon. Wea. Rev.*, **126**, 1397–1418, [https://doi.org/10.1175/1520-0493\(1998\)126<1397:TOCMGE>2.0.CO;2](https://doi.org/10.1175/1520-0493(1998)126<1397:TOCMGE>2.0.CO;2).
- , S. Gravel, A. Méthot, A. Patoine, M. Roch, and A. Staniforth, 1998b: The operational CMC-MRB Global Environmental Multiscale (GEM) model. Part I: Design considerations and formulation. *Mon. Wea. Rev.*, **126**, 1373–1395, [https://doi.org/10.1175/1520-0493\(1998\)126<1373:TOCMGE>2.0.CO;2](https://doi.org/10.1175/1520-0493(1998)126<1373:TOCMGE>2.0.CO;2).
- Delage, Y., 1997: Parameterising sub-grid scale vertical transport in atmospheric models under statically stable conditions. *Bound.-Layer Meteor.*, **82**, 23–48, <https://doi.org/10.1023/A:1000132524077>.
- Díaz-Isaac, L. I., T. Lauvaux, M. Bocquet, and K. J. Davis, 2019: Calibration of a multi-physics ensemble for estimating the uncertainty of a greenhouse gas atmospheric transport model. *Atmos. Chem. Phys.*, **19**, 5695–5718, <https://doi.org/10.5194/acp-19-5695-2019>.
- Girard, C., and Coauthors, 2014: Staggered vertical discretization of the Canadian Environmental Multiscale (GEM) model using a coordinate of the log-hydrostatic-pressure type. *Mon. Wea. Rev.*, **142**, 1183–1196, <https://doi.org/10.1175/MWR-D-13-00255.1>.
- Gneiting, T., and A. Raftery, 2007: Strictly proper scoring rules, prediction and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, <https://doi.org/10.1198/016214506000001437>.
- Gordon, N. J., D. J. Salmond, and A. F. M. Smith, 1993: Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEEE Proc.*, **140**, 107–113, <https://doi.org/10.1049/ip-f-2.1993.0015>.
- Grell, G. A., and D. Dévényi, 2002: A generalized approach to parameterizing convection combining ensemble and data assimilation techniques. *Geophys. Res. Lett.*, **29**, 1693, <https://doi.org/10.1029/2002GL015311>.
- Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, [https://doi.org/10.1175/1520-0434\(2000\)015<0559:DOTCRP>2.0.CO;2](https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2).
- Houtekamer, P. L., 2011: The use of multiple parameterizations in ensembles. *Proc. ECMWF Workshop on Representing Model Uncertainty and Error in Numerical Weather and Climate Prediction Models*, Shinfield Park, Reading, ECMWF, 163–173.
- , and L. Leflaivre, 1997: Using ensemble forecasts for model validation. *Mon. Wea. Rev.*, **125**, 2416–2426, [https://doi.org/10.1175/1520-0493\(1997\)125<2416:UEFFMV>2.0.CO;2](https://doi.org/10.1175/1520-0493(1997)125<2416:UEFFMV>2.0.CO;2).
- , and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.*, **126**, 796–811, [https://doi.org/10.1175/1520-0493\(1998\)126<0796:DAUAEK>2.0.CO;2](https://doi.org/10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2).
- , L. Leflaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.*, **124**, 1225–1242, [https://doi.org/10.1175/1520-0493\(1996\)124<1225:ASSATE>2.0.CO;2](https://doi.org/10.1175/1520-0493(1996)124<1225:ASSATE>2.0.CO;2).
- , H. L. Mitchell, and X. Deng, 2009: Model error representation in an operational ensemble Kalman filter. *Mon. Wea. Rev.*, **137**, 2126–2143, <https://doi.org/10.1175/2008MWR2737.1>.
- , M. Buehner, and M. De La Chevrotière, 2018: Using the hybrid gain algorithm to sample data assimilation uncertainty. *Quart. J. Roy. Meteor. Soc.*, **145**, 35–56, <https://doi.org/10.1002/QJ.3426>.
- Huffman, G., D. Bolvin, D. Braithwaite, K. Hsu, R. Joyce, and P. Xie, 2014: Integrated multi-satellite retrievals for GPM (IMERG), version 4.4. NASA's precipitation processing center, accessed 25 February, 2020, <https://doi.org/10.5067/GPM/IMERG/3B-HH/05>.
- Järvinen, H., M. Laine, A. Solonen, and H. Haario, 2012: Ensemble prediction and parameter estimation system: The concept. *Quart. J. Roy. Meteor. Soc.*, **138**, 281–288, <https://doi.org/10.1002/qj.923>.
- Kain, J. S., and J. M. Fritsch, 1990: A one-dimensional entraining/detraining plume model and its application in convective parameterization. *J. Atmos. Sci.*, **47**, 2784–2802, [https://doi.org/10.1175/1520-0469\(1990\)047<2784:AODEPM>2.0.CO;2](https://doi.org/10.1175/1520-0469(1990)047<2784:AODEPM>2.0.CO;2).
- , and —, 1992: The role of the convective “trigger function” in numerical forecasts of mesoscale convective systems. *Meteor. Atmos. Phys.*, **49**, 93–106, <https://doi.org/10.1007/BF01025402>.
- Laine, M., A. Solonen, H. Haario, and H. Järvinen, 2012: Ensemble prediction and parameter estimation system: The method. *Quart. J. Roy. Meteor. Soc.*, **138**, 289–297, <https://doi.org/10.1002/qj.922>.
- Lott, F., and M. J. Miller, 1997: A new subgrid-scale orographic drag parameterization: Its formulation and testing. *Quart. J. Roy. Meteor. Soc.*, **123**, 101–127, <https://doi.org/10.1002/qj.49712353704>.
- McTaggart-Cowan, R., P. Vaillancourt, A. Zadra, L. Separovic, S. Corvec, and D. Kirshbaum, 2019a: A Lagrangian perspective on parameterizing deep convection. *Mon. Wea. Rev.*, **147**, 4127–4149, <https://doi.org/10.1175/MWR-D-19-0164.1>.
- , and Coauthors, 2019b: Modernization of atmospheric physics parameterization in Canadian NWP. *J. Adv. Model. Earth Syst.*, **11**, 3593–3635, <https://doi.org/10.1029/2019MS001781>.
- Ollinaho, P., M. Laine, A. Solonen, H. Haario, and H. Järvinen, 2013: NWP model forecast skill optimization via closure parameter variations. *Quart. J. Roy. Meteor. Soc.*, **139**, 1520–1532, <https://doi.org/10.1002/qj.2044>.
- Orrell, D., L. Smith, J. Barkmeijer, and T. N. Palmer, 2001: Model error in weather forecasting. *Nonlinear Processes Geophys.*, **8**, 357–371, <https://doi.org/10.5194/npg-8-357-2001>.
- Penny, S. G., 2014: The hybrid local ensemble transform Kalman filter. *Mon. Wea. Rev.*, **142**, 2139–2149, <https://doi.org/10.1175/MWR-D-13-00131.1>.
- Ruiz, J. J., M. Pulido, and T. Miyoshi, 2013: Estimating model parameters with ensemble-based data assimilation: A review. *J. Meteor. Soc. Japan*, **91**, 79–99, <https://doi.org/10.2151/jmsj.2013-201>.
- Schirber, S., D. Klocke, R. Pincus, J. Quaas, and J. Anderson, 2013: Parameter estimation using data assimilation in an

- atmospheric general circulation model: From a perfect toward the real world. *J. Adv. Model. Earth Syst.*, **5**, 58–70, <https://doi.org/10.1029/2012MS000167>.
- Snyder, C., T. Bengtsson, P. Bickel, and J. Anderson, 2008: Obstacles to high-dimensional particle filtering. *Mon. Wea. Rev.*, **136**, 4629–4640, <https://doi.org/10.1175/2008MWR2529.1>.
- Trenberth, K., and C. Guillemot, 1998: Evaluation of the atmospheric moisture and hydrological cycle in the NCEP/NCAR reanalyses. *Climate Dyn.*, **14**, 213–231, <https://doi.org/10.1007/s003820050219>.
- Wilks, D., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. International Geophysics Series, Vol. 100, Academic Press, 648 pp.
- Zamo, M., and P. Naveau, 2018: Estimation of the continuous ranked probability score with limited information and applications to ensemble weather forecasts. *Math. Geosci.*, **50**, 209–234, <https://doi.org/10.1007/s11004-017-9709-7>.
- Zhu, Y., and Coauthors, 2016: All-sky microwave radiance assimilation in NCEP's GSI analysis system. *Mon. Wea. Rev.*, **144**, 4709–4735, <https://doi.org/10.1175/MWR-D-15-0445.1>.