Nonlin. Processes Geophys., 20, 1001–1010, 2013
www.nonlin-processes-geophys.net/20/1001/2013/
doi:10.5194/npg-20-1001-2013

Nonlinear Processes in Geophysics
Open Access

Parameter variations in prediction skill optimization at ECMWF

P. Ollinaho[1,2], P. Bechtold[3], M. Leutbecher[3], M. Laine[1], A. Solonen[1,4], H. Haario[1,4], and H. Järvinen[2]

[1]Finnish Meteorological Institute, Erik Palménin aukio 1, Helsinki, Finland
[2]University of Helsinki, Department of Physics, Gustaf Hällströmin katu 2a, Helsinki, Finland
[3]European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, UK
[4]Lappeenranta University of Technology, Skinnarilankatu 34, Lappeenranta, Finland

Correspondence to: P. Ollinaho (pirkka.ollinaho@fmi.fi)

Received: 11 March 2013 – Revised: 23 July 2013 – Accepted: 21 October 2013 – Published: 22 November 2013

Abstract. Algorithmic numerical weather prediction (NWP) skill optimization has been tested using the Integrated Forecasting System (IFS) of the European Centre for Medium-Range Weather Forecasts (ECMWF). We report the results of initial experimentation using importance sampling based on model parameter estimation methodology targeted for ensemble prediction systems, called the ensemble prediction and parameter estimation system (EPPES). The same methodology was earlier proven to be a viable concept in low-order ordinary differential equation systems, and in large-scale atmospheric general circulation models (ECHAM5). Here we show that prediction skill optimization is possible even in the context of a system that is (i) of very high dimensionality, and (ii) carefully tuned to very high skill. We concentrate on four closure parameters related to the parameterizations of sub-grid scale physical processes of convection and formation of convective precipitation. We launch standard ensembles of medium-range predictions such that each member uses different values of the four parameters, and make sequential statistical inferences about the parameter values. Our target criterion is the squared forecast error of the 500 hPa geopotential height at day three and day ten. The EPPES methodology is able to converge towards closure parameter values that optimize the target criterion. Therefore, we conclude that estimation and cost function-based tuning of low-dimensional static model parameters is possible despite the very high dimensional state space, as well as the presence of stochastic noise due to initial state and physical tendency perturbations. The remaining question before EPPES can be considered as a generally applicable tool in model development is the correct formulation of the target criterion. The one used here is, in our view, very selective. Considering the multi-faceted question of improving forecast model performance, a more general target criterion should be developed. This is a topic of ongoing research.

1 Introduction

Long-term improvements in numerical weather prediction models (NWP) originate from dedicated research to improve the representation of atmospheric phenomena across all spatial and temporal scales. This involves a slow but steady development process that gradually improves the predictive skill of NWP models and reduces their systematic errors (Simmons and Hollingsworth, 2002). The increased operational skill can be attributed to improvements in all prediction system components over many prediction system generations, and covers observing systems, data assimilation, forecast models, and high-performance computing capabilities. Current thinking is that this gradual progress of the past decades will continue into the future.

Short-term prospects for prediction skill improvements are quite different. Short-term developments are typically incremental, such as refinements to existing modeling schemes, or the introduction of new observing system components. These are aimed to be implemented as new model releases within a time frame of some months and are seen as gradual small steps between model generations. For instance, parameterization schemes of sub-grid scale physical processes typically undergo many refinements during their lifetime, while entire modules of physical processes are replaced relatively infrequently. It is a generally accepted fact that in forecast systems tuned to high predictive skill, the introduction of new and more physically justified schemes seldom leads to skill improvements without careful and time-consuming model retuning. In this respect, tunable model parameters provide a practical way to modify the model behavior and tune the skill, since model resolution, parameterization paradigm, and other structural matters are usually fixed.

In order to facilitate the model re-tuning, some algorithmic tools would be advantageous to save time and effort and speed up operational implementations. Moreover, in research, the model code is typically modified frequently as new ideas are tested. It is commonplace that these research tests are inconclusive, because in the modified modeling system various multi-scale interactions and dynamics-physics feedbacks are not tuned into harmony. These considerations motivate the search for simple-to-use and accurate yet computationally affordable model tuning algorithms. At the same time one has to acknowledge that re-tuning of complex multiscale modeling systems by optimizing closure parameter values is an extremely hard problem, and there are almost certainly no simple solutions available. The basic reason for this is the fact that while Navier–Stokes systems tend to “forget” the initial values, the impact of parameter values accumulate to the state variables with time, and thus this constitutes a particularly sensitive inverse problem. Therefore, even a partial solution to the problem would be beneficial. Such a solution would be, for instance, a method to provide re-tuned “candidate” models that would then be passed for closer inspection from various aspects. Even this would be a step forward from the current predominantly trial-and-error procedures.

In this paper we will continue to study an ensemble-based method to estimate optimal closure parameter values and their uncertainties. The Ensemble Prediction and Parameter Estimation System (EPPES; Järvinen et al., 2012; Laine et al., 2012) utilizes ensemble prediction systems to make statistical inferences about the NWP model closure parameters as follows. A set of model closure parameters is selected, and its prior probability distribution is specified based on expert knowledge as a Gaussian with the distribution parameters being the mean and standard deviation. A sample is drawn from this distribution so that each ensemble member has different parameter values that do not change during the integration. Once observations are available, a likelihood function is evaluated for each member, and parameter values are weighted according to their likelihood. Re-sample from the likelihood-weighted prior is, in fact, a sample from the posterior distribution of the parameter. Such a re-sample is used to update the prior distribution parameters (mean and standard deviation). The parameter estimation proceeds sequentially as the prior parameter distribution for the current ensemble is first updated to become a posterior distribution, which is then used as a prior distribution for the next ensemble.

The approach has been shown to perform as intended in low-order systems, as well as in atmospheric general circulation model ECHAM5 (Roeckner et al., 2003) at low resolution (Ollinaho et al., 2013). The main remaining questions are as follows: (i) are the convergence properties of the EPPES algorithm in the low-dimensional parameter space preserved as the model state space becomes very high-dimensional, (ii) do the stochastic model physics perturbations affect the estimation process detrimentally, and (iii) is it possible to formulate a target criterion (likelihood function) such that the parameter estimation results in a genuine and universally acceptable model improvement? This paper explores questions (i) and (ii), while question (iii) remains a topic for further research and is only briefly discussed here.

In this paper we present experimentation using the European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System (IFS), including their Ensemble Prediction System (EPS). The experimental setup is thus close to an operational system, but not quite identical, since the forecast model resolution is lower than in the operational system. However, several aspects are now more realistic than in our earlier experimentation using the ECHAM5 climate model (Ollinaho et al., 2013). The forecast model resolution has been increased from triangular truncation 42 and 31 vertical levels (T42L31) in ECHAM5 to $\mathrm{T_L}159\mathrm{L}62$ in the IFS forecast model. The EPS is now a genuine system with “native” initial state perturbations and model uncertainty representation, in contrast to the earlier “EPS emulator” in the context of ECHAM5. Finally, and perhaps most importantly, the IFS forecast model is tuned to a very high level of forecast skill, and therefore it is certainly very hard to gain any further skill improvements. The ECHAM5 model, although a very good climate model, was not tuned to skilful medium-range weather forecasting. This may partly explain the good performance of the EPPES algorithm, as reported in Ollinaho et al. (2013). We present the experimental setup in Sect. 2, the parameter estimation and validation results in Sect. 3, before the Discussion and Conclusions.

2 Experimental setup

2.1 The IFS model and subset of parameters

In the experimentation, we use the IFS version that was operational from November 2011 to June 2012 (CY37R3)[1], but at a lower resolution. The forecast model of the IFS is a global hydrostatic general circulation model of the atmosphere with a spectral, semi-implicit, and semi-Lagrangian two time-level dynamical solver. We use the model at spectral truncation $\mathrm{T_L}159$ (about 125 km) with 62 vertical levels and the model top at 5 hPa. The time step for the model dynamics and physical parameterizations is 30 min, with the exception of radiative transfer, which is calculated once every 3 h. The model contains a range of parameterizations for physical processes with their specific closure schemes. The

[1] IFS documentation is available online at http://www.ecmwf.int/research/ifsdocs.

Table 1. The sub-set of IFS closure parameters with time-invariant parameter variations.

Parameter	Description
ENTRORG	Entrainment rate for positively buoyant deep convection
ENSHALP	Shallow entrainment defined as ENTSHALP × ENTRORG
DETRPEN	Detrainment rate for penetrative convection
RPRCON	Coefficient for determining conversion from cloud water to rain

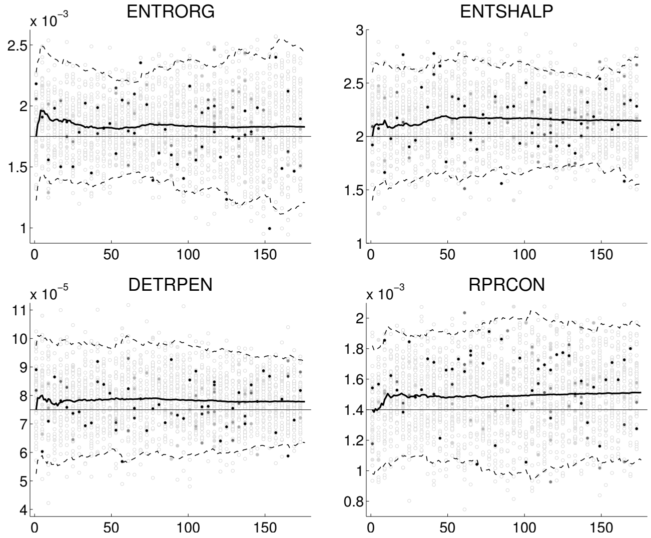

Fig. 1. Time evolution of the parameter values in 177 consecutive ensembles. A vertical column of markers represents parameter values of one ensemble. The darker colors correspond to values with high likelihood. The parameter distribution mean value $\boldsymbol{\mu}$ (thick line) and $\boldsymbol{\mu} \pm 2 \times$ standard deviation (dashed lines) are also shown. For clarity, the default parameter value (thin horizontal line), and every fourth ensemble only is plotted.

experiments reported here concern the estimation of parametric uncertainties of convection. It is represented by a bulk mass flux scheme (Tiedtke, 1989; Bechtold et al., 2008), and further divided into deep, mid-level, and shallow convection. The formation of convective precipitation is determined by the conversion rate from cloud water into rain, evaporation of precipitation, and the melting rate of snow. In contrast to the original Tiedtke (1989) scheme as used earlier in the ECHAM5 runs (Ollinaho et al., 2013), the entrainment and detrainment formulation in the IFS is shown to closely follow observations and data from cloud-resolving models (de Rooy, 2013). This makes it even more difficult in practice to further improve on these parameters here.

The optimization of prediction skill here involves four prediction model closure parameters related to entrainment and detrainment rates in deep convection, entrainment in shallow convection, and precipitation formation (Table 1). The choice of these particular parameters is motivated as follows. First, the set of parameters has to be rather small for the estimation to converge with affordable sampling. In our previous experimentation with the ECHAM5 climate model, four and seven parameters were successfully varied simultaneously. Second, expert knowledge supports this choice of parameters. Individually, they are known to affect mostly the tropical troposphere. One has to bear in mind, however, that individual impacts due to the parameter variations are based on sensitivity studies, but the system response to the joint variation of all parameters is much less explored. Finally, the parameters in the experiments with the ECHAM5 climate model were very similar to the ones in Table 1, and thus we can concentrate here on the impacts of increasing resolution and more realistic stochastic physics on the estimation task.

2.2 The ensemble prediction system

Initial state perturbations in the Centre's ensemble prediction systems combine two sources. A lower resolution ensemble of data assimilations (EDA) is run in parallel to high-resolution data assimilation. The ensemble of background states is used to generate the initial perturbations. These are complemented by perturbations based on initial-time singular vectors (Buizza et al., 2008; Isaksen et al., 2010). Uncertainty of the forecast model formulation is represented in these experiments by stochastically perturbing the tendencies generated by the parameterization schemes (Buizza et al., 1999; Palmer et al., 2009) and by a stochastic kinetic energy backscatter scheme that adds a stream function forcing to the momentum equation (Berner et al., 2009).

2.3 Implementation of the estimation algorithm

Details of the ensemble prediction and parameter estimation system (EPPES) can be found in Laine et al. (2012), which applied the algorithm to a modified Lorenz-95 system (Lorenz, 1995; Wilks, 2005). The implementation here follows closely the one presented in Ollinaho et al. (2013), which used an EPS emulator. Thus only an outline is provided here.

In EPPES, it is assumed that for time window i, the optimal model parameter $\boldsymbol{\theta}_i$ is a realization of a p-dimensional random vector, for which we assume a multivariate Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix Σ

$$\boldsymbol{\theta}_i \sim N(\boldsymbol{\mu}, \Sigma), i = 1, 2, \ldots$$

Table 2. IFS parameter values applied in the EPPES tests. Prior mean values correspond to the default model values. Prior standard deviation (the standard deviation of the proposal distribution of the first ensemble) and bounds (minimum and maximum allowed parameter values) are subjectively specified. Posterior mean and standard deviation are the EPPES estimates after 177 estimation steps with the specified cost function.

Parameter	Prior		Bounds	Posterior	
	mean	std. dev.		mean	std. dev.
ENTRORG	1.75×10^{-3}	2.63×10^{-4}	$0.9–2.6 \times 10^{-3}$	1.83×10^{-3}	3.11×10^{-4}
ENSHALP	2.00	0.30	1.0–3.0	2.15	0.30
DETRPEN	0.75×10^{-4}	1.13×10^{-5}	$0.4–1.1 \times 10^{-4}$	0.78×10^{-4}	0.72×10^{-5}
RPRCON	1.40×10^{-3}	2.10×10^{-4}	$0.7–2.1 \times 10^{-3}$	1.51×10^{-3}	2.22×10^{-4}

The distribution parameters $\boldsymbol{\mu}$ and Σ are assumed to be unknown but static in time. In EPPES, the problem of estimating the model parameter $\boldsymbol{\theta}$ is formulated as a problem of estimating the distribution parameters (or, hyperparameters) $\boldsymbol{\mu}$ and Σ. The interpretation is that there is a mean parameter value $\boldsymbol{\mu}$ that performs best on average considering all weather types, seasons, etc., but due to the evident modeling errors, the optimal parameter value varies according to Σ in different time windows. Here, the dimension p of the parameter vector equals 4.

EPPES is closely related to other ensemble-based estimation methods, such as the particle filter (Kivman, 2003; van Leeuwen, 2003). It is based on importance-sampling ideas. Instead of considering the parameter sample as particles that are propagated in time, they are re-sampled each time from an updated parameterized parameter perturbation proposal distribution. This way the well-known problem of collapse of weights in particle filters does not have a deteriorating effect on the estimation.

Instead of estimating the actual parameter $\boldsymbol{\theta}$, we aim for the middle time window variability of locally optimal $\boldsymbol{\theta}$. This is achieved using hierarchical formulation of uncertainties with hyperparameters $\boldsymbol{\mu}$ and Σ. The fundamental idea behind EPPES is that only these hyperparameters related to the proposal distribution are updated. This allows us to circumvent many problems encountered in the estimation of static model parameters in data assimilation frameworks (see, e.g., Rougier, 2013).

Initially, the parameters $\boldsymbol{\mu}$ and Σ are specified according to expert knowledge ("prior" in Table 2) with a diagonal covariance Σ, i.e., no prior knowledge about the parameter covariance is assumed. Because a Gaussian distribution is used, parameter bounds are set to prevent the occurrence of nonphysical parameter values (Table 2). Then, a sample is drawn from this prior distribution, and an ensemble of predictions is generated using these parameters values. The likelihood of each prediction is then evaluated as a fit to analyses, and each parameter vector is weighted by the likelihood. A re-sample is drawn from the weighted parameter sample, which favors well-performing parameter values associated with high likelihood. In statistics, this mechanism is known as importance sampling, and the re-sampled values can be considered as samples from the posterior distributions. Now, the weighted sample is used to update the hyperparameters $\boldsymbol{\mu}$ and Σ. The covariance matrix Σ represents the middle ensemble variability of the parameter vector $\boldsymbol{\theta}$ around the mean parameter $\boldsymbol{\mu}$.

In the experiments with the IFS, the cost function is formulated as a sum of three and ten day squared forecast errors as follows:

$$J(\boldsymbol{\theta}) = 10 \cdot \sum_A \left(z_{\mathrm{f}}^{72}(\boldsymbol{\theta}) - z_{\mathrm{a}}\right)^2 \mathrm{d}A + \sum_A \left(z_{\mathrm{f}}^{240}(\boldsymbol{\theta}) - z_{\mathrm{a}}\right)^2 \mathrm{d}A.$$

Here z_{f}^{72} (z_{f}^{240}) is a 72 h (240 h) forecast of the 500 hPa geopotential height, z_{a} the verifying operational analysis of ECMWF valid at the 72 and 240 h forecast ranges, respectively, and dA the areal element of the model grid. The factor 10 makes the two right-hand terms approximately equal in magnitude, and to some extent balances their contributions to the cost function. The parameters $\boldsymbol{\theta}$ in the formula imply that the forecasts depend on the sampled parameter values. We note that the cost function is closely related to the root-mean-squared forecast error (RMSE) commonly used as a validation metric in NWP. Finally, the likelihood is defined as $\exp(-1/2J(\boldsymbol{\theta}))$. Note that EPPES as such requires very little additional computing time, as it essentially monitors the computations of an EPS system.

2.4 The experiments

The experiment (referred to as "ParVar") consists of running a sequence of 50 member ensembles with initial-state perturbations, and applying initial time parameter variations. In addition, a control member is run for each ensemble without initial perturbations, and with default parameter values; this member does not affect the parameter distribution update. The period of 12 May 2011 to 8 August 2011 was covered twice a day (00:00 and 12:00 UTC). Thus, 177 ensembles were generated, equaling 8850 test forecasts with different parameter combinations. Moreover, an ensemble without parameter perturbations has been run as a reference. It is referred to as "Ctrl" and will be discussed in Sect. 3.3.

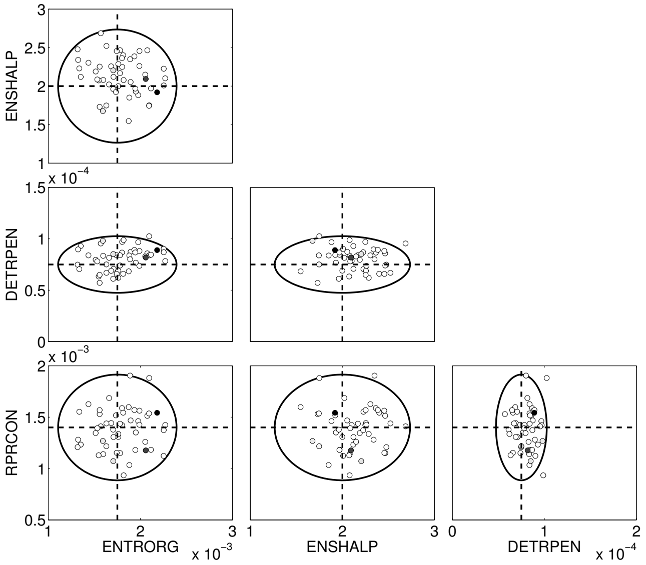

Fig. 2a. Pair-wise parameter covariances at the initial time. Default parameter values ($\boldsymbol{\mu}_0$) are denoted by dashed lines. The ellipse represents the prior parameter uncertainty as specified initially (the 95 % probability range of the parameter uncertainty Σ_0). The small markers are the proposed parameter values at the first step; darkness of color is indicative of the weights given to re-sampled parameter values.

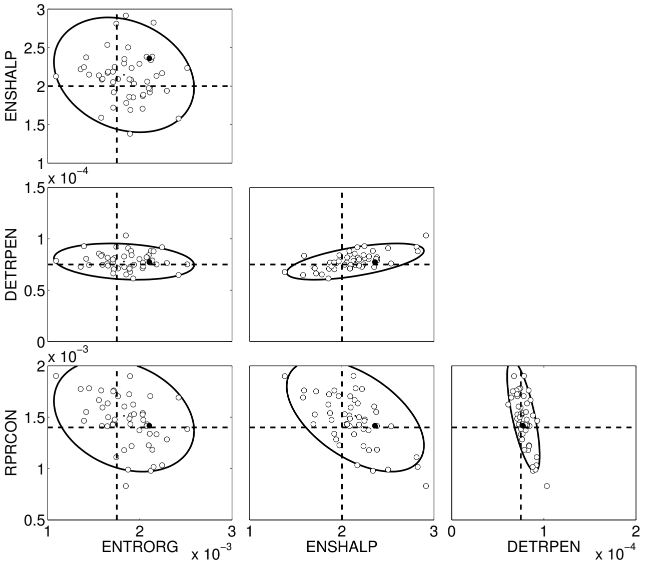

Fig. 2b. As Fig. 2a, but after 177 consecutive ensembles; the small markers are the proposed parameter values at step 177.

3 Results

3.1 Evolution of parameter distributions

The evolution of the four parameter values in the 177 consecutive ensembles is given in Fig. 1. A vertical column of markers represents parameter values of one ensemble. Dark markers correspond to parameter values with high likelihood. The parameter distribution mean value $\boldsymbol{\mu}$ (thick line) changes conservatively after the initial "shock", and remains above the default parameter value (thin horizontal line) by 4–8 % for all four parameters. Note, for instance, that the dark markers for RPRCON are mostly above the default parameter value, thus "pulling" the mean upwards. The square roots of the diagonal of the distribution parameter Σ give the distribution standard deviations, shown in Fig. 1 as $\boldsymbol{\mu} \pm 2\times$ standard deviation (dashed lines). It reduces markedly (about 36 %) for DETRPEN, while for other parameters it increases. The final distribution mean and standard deviations are shown in Table 2 as posterior values.

The parameter pair-wise covariance ellipses, each corresponding to the 95 % probability region, are presented in Fig. 2 at the initial time (Fig. 2a), and after 177 estimation steps (Fig. 2b). Initially (Fig. 2a), the model parameters are assumed to be independent, and the specified prior parameter uncertainties appear as ellipses centered at the default value $\boldsymbol{\mu}_0$ (dashed lines). The small markers denote the sample drawn from the prior distribution. After 177 sampling steps (Fig. 2b), the covariance ellipses appear at the new distribution mean values $\boldsymbol{\mu}$, and some are tilted (for instance, DETRPEN vs. RPRCON). This indicates that these parameters are mutually correlated. The mutual correlation coefficients evolve more slowly than the mean values (not shown). They converge gradually towards their final values mainly during the first 100 estimation steps. For instance, the strongest correlations are −0.7 for between DETRPEN and RPRCON, and +0.6 between ENSHALP and DETRPEN. They reach values −0.4 (+0.4) already after 55 (40) iterations.

Note that the default parameter values are inside the posterior 95 % confidence range (Fig. 2b). This is indicative of the accurate tuning of the default IFS model, and is in contrast to the experiments with the ECHAM5 model (Ollinaho et al., 2013).

3.2 Validation of the optimized model

The experiment is validated by running the model with the default and posterior mean parameter values (Table 2) for the period 12 May to 8 August 2011. Note that this is the same period as used for the parameter estimation. A 10-day forecast is launched every 48 h at 00:00 UTC, totaling 45 forecasts. Initial states for the forecasts are the operational analyses of the ECMWF without re-doing data assimilation. The additional effects of the posterior parameter values via data assimilation are thus ignored. Also, forecast verification makes use of the ECMWF operational analyses.

We first check that the cost function is smaller in the optimized than in the default model, which is the necessary condition for the estimation procedure to deliver. In the validation set of 45 forecasts, the cost function is indeed reduced. However, only the 72 h forecast error contribution separately (i.e., the first term of the cost function) is reduced at the 95 % confidence level. We consider this condition satisfied and now proceed to a more detailed validation.

The posterior parameters of Table 2 are validated in forecast experiments. Next, three metrics of the 500 hPa geopotential height are used: mean error, root-mean-squared forecast error (RMSE), and anomaly correlation coefficient (ACC), defined as

$$\mathrm{ACC} = \frac{\sum \mathrm{d}z_\mathrm{f}\mathrm{d}z_\mathrm{a}}{\left(\sum (\mathrm{d}z_\mathrm{f})^2 \sum (\mathrm{d}z_\mathrm{a})^2\right)^{\frac{1}{2}}}.$$

Here $\mathrm{d}z_\mathrm{f}$ and $\mathrm{d}z_\mathrm{a}$ are the forecast and analysis anomalies with respect to the climatological mean, which depends on the day of the year and location. These two metrics complement each other, since RMSE penalizes forecast bias, while ACC penalizes incorrect patterns in forecast fields. Thus, if RMSE is decreased while ACC is not significantly degraded, we can conclude that the skill improvement is not due to smoothing effects, but related either to bias reduction and/or more accurate forecasts of spatial variations in the height field. Note that the optimization criterion (likelihood) is closely related to RMSE, while ACC is more independent of the criterion used in the estimation.

The optimized model parameters have their largest impact on forecasts in the tropics. Thus the validation results at 500 hPa up to a 10-day forecast range are presented first for the latitude band 20° S to 20° N. Figure 3 shows the forecast skill differences between the default and optimized model for the three metrics. The notation is such that a positive difference implies that the optimized model is more accurate than the default model. In Fig. 3a, the mean error is positive up to day 6 for all individual forecasts (dots). The mean over all cases (continuous line) remains positive throughout the 10-day range. The 95 % confidence interval of the mean (vertical bars) first meets the zero line at day 9.5. The RMSE is qualitatively similar to the mean error. In Fig. 3b, the RMSE is positive up to day 4.5 for all individual forecasts (dots). The mean over all cases (continuous line) remains positive throughout the 10-day range. The 95 % confidence interval of the mean of RMSE first meets the zero line at day 10. The ACC is generally positive as well. In Fig. 3c, the mean over all cases (continuous line) is positive throughout the 10-day range, except that at day 8.5 it touches the zero line. The 95 % confidence interval of the mean ACC first meets the zero line at day 4.5.

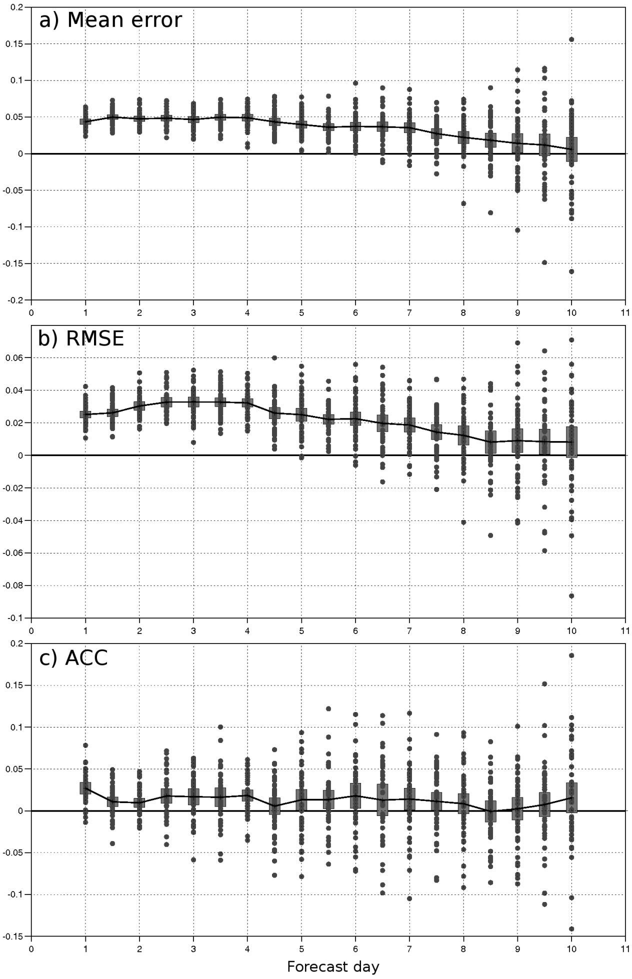

Fig. 3. Forecast skill score differences between the default model and the optimized model for the 500 hPa geopotential height in the tropics (20° S to 20° N). Notation: positive difference implies that the optimized model is more accurate. **(a)** Mean error, **(b)** RMSE, **(c)** ACC. Included are 45 forecast cases between 12 May and 8 August 2012 for individual score difference (dots), its mean (continuous line) and the 95 % confidence interval of the mean (vertical bars).

Next, a comprehensive set of forecast verification results is presented using a so-called scorecard (Fig. 4). It is a concise presentation of a large number of scores for various geographical regions, variables, levels, and forecast ranges. In total, the scorecard contains 1710 individual scoring elements. The notation is such that green (red) colors indicate the optimized model scoring better (worse) that the default model. Small and large arrow heads up (down) indicate that the result is significant at 95 % or 99 % confidence level, respectively, for the optimized (default) model to score better. White boxes indicate the models performing equally well.

The main features in Fig. 4 are as follows. First, there is striking 99 % significant global degradation of the 100 hPa geopotential height RMSE. This feature can be explained as follows. The likelihood formulation targets the forecast error of the 500 hPa geopotential height, and indeed the optimized model has a significantly reduced RMSE and mean error at 500 hPa geopotential (as seen in Fig. 4 in the tropics, and in Fig. 3a). The side effect is that the improved 500 hPa height has been reached at the expense of geopotential height at higher levels (at 100 hPa, and very likely also at 200 and

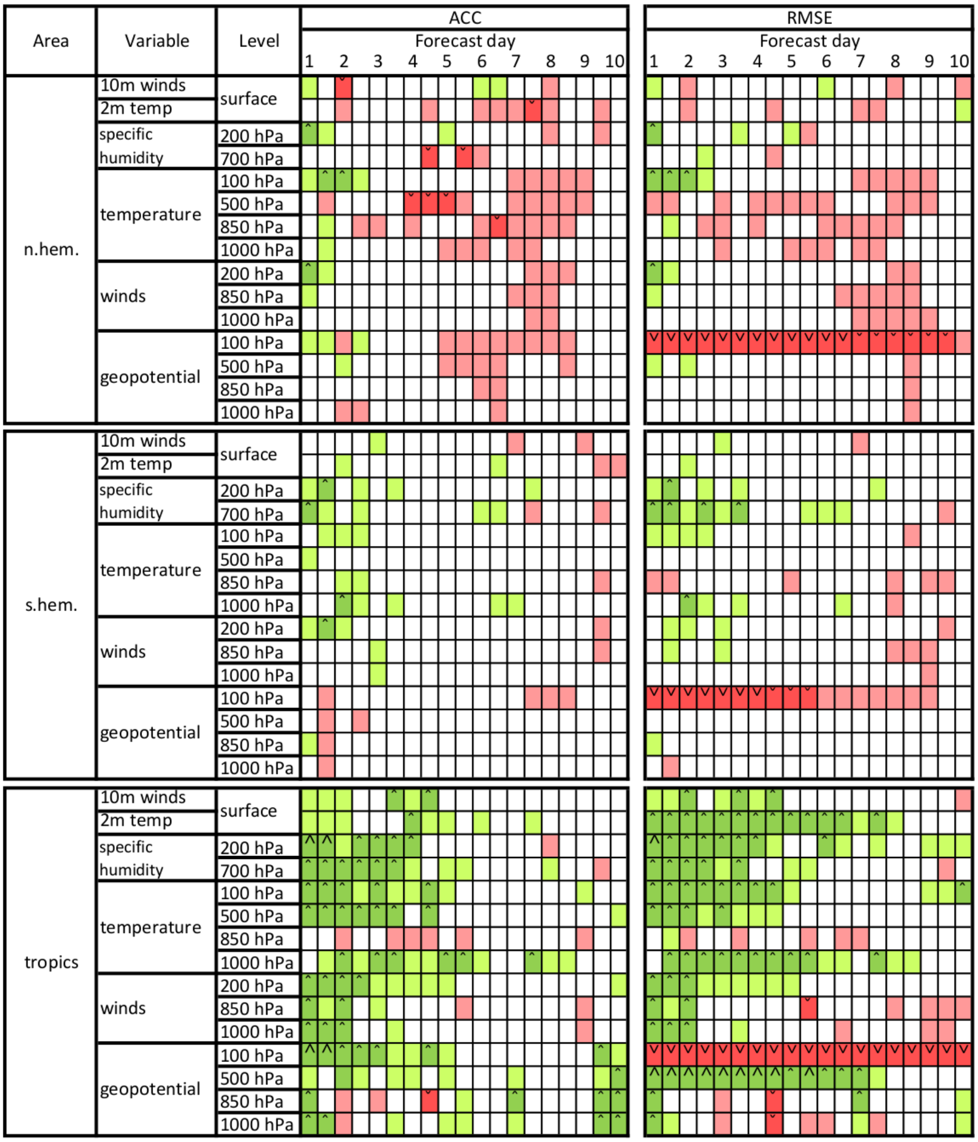

Fig. 4. A forecast validation scorecard for the 45 forecast cases between 12 May 2012 and 8 August 2012 using the following color code: green is good for the optimized model, while red is good for the default model. Small (large) arrow head indicates 95 % (99 %) level of statistical significance of the sore difference. The 1st column indicates the area, the 2nd the variable, the 3rd pressure level, and the 4th and 5th columns the ACC and RMS score for forecast days 1–10.

50 hPa). Note, however, that the corresponding ACC is significantly improved in the short-range predictions, thus implying that the RMSE degradation is due to increased bias rather than incorrect height patterns. Second, there is a remarkable tropical score improvement for temperature and humidity up to about day 5, and winds up to about day 2. In fact, apart from the degraded 100 hPa height RMSE, the tropics benefit considerably from the modified parameter values. The improvement in the winds is especially impressive, as it is a very important variable in the tropical troposphere, and it is generally very hard to improve wind scores in that region. Convection also plays an important role in the mid-latitude storm tracks. The effects of the convection parameter changes can thus be seen in the middle latitude height and wind scores. While these scores are positive in the Southern Hemisphere in the short range, there is some degradation in the Northern Hemisphere scores in the medium range.

3.3 Impact on the ensemble prediction system

The parameter perturbations cause additional ensemble spread on top of the dispersion due to initial condition perturbations and stochastic physics perturbations. Although the main purpose of the parameter perturbations generated from

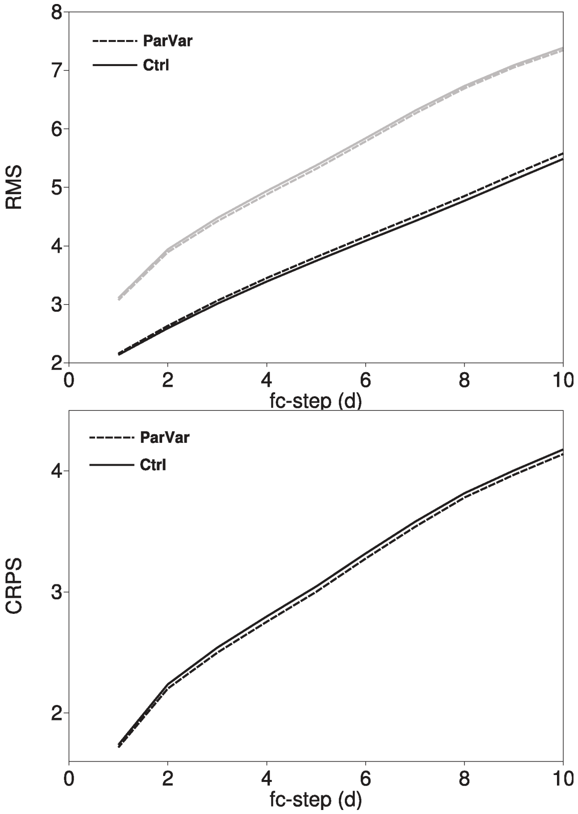

Fig. 5. Ensemble verification of the ensemble with parameter variations (ParVar) and a control ensemble (Ctrl) that uses the same initial perturbations and model uncertainty representation but no parameter perturbations for the 200 hPa zonal wind component in the tropics: **(a)** ensemble standard deviation (black) and ensemble mean RMS error (grey); **(b)** Continuous Ranked Probability Score. Sample of 90 cases in the period 24 June, 12:00 UTC to 8 August, 00:00 UTC.

the EPPES algorithm is to sample the parameter space and test the model response, they provide an additional representation for model uncertainties. No changes to either initial perturbations or the stochastic physics schemes were made in order to improve the spread–error relationship at any stage of the experimentation.

Now, we examine the impact of the parameter variations on the ensemble forecasts. A control ensemble (Ctrl) serves as a reference that uses the default values of the four parameters for all members. Otherwise, the ensemble configuration of experiment Ctrl is identical to the experiment with parameter variations (ParVar). In order to omit the initial phase during which the parameter distribution still evolves more rapidly, verification statistics have been averaged for the last 90 ensemble forecasts only. This covers the period from 24 June, 12:00 UTC to 8 August, 00:00 UTC.

The parameter variations generate additional ensemble variance mostly in tropical regions. Figure 5a shows the ensemble standard deviation and the ensemble mean RMS error for the 200 hPa zonal wind component in the tropics. For both experiments, the ensemble standard deviation is smaller than the ensemble mean RMS error. Due to the lower horizontal resolution, both ensembles are more underdispersive than the operational ensemble configuration, which has a horizontal resolution of $\mathrm{T_L}639$. Experiment ParVar has more spread and a lower ensemble mean RMS error than experiment Ctrl.

The probabilistic skill of the two ensemble experiments is quantified with the Continuous Ranked Probability Score (CRPS). The CRPS for 200 hPa zonal wind in the tropics is shown in Fig. 5b. Experiment ParVar is generally more skilful than experiment Ctrl in the tropics, except for temperature around 200 hPa (not shown). The impact on CRPS in the extra-tropics is close to neutral (not shown). The improvement that is observed in ParVar may be due to two aspects. First, the reliability has been improved as the ensemble spread better matches the RMS error of the ensemble mean. Secondly, the average skill of the ensemble members in ParVar is higher than in Ctrl as the mean of the parameter distribution ($\boldsymbol{\mu}$) has changed. The parameter covariance (Σ) guides the parameter sampling towards the well-performing ones, too. It is left for future work to determine whether one of the two aspects dominates the skill improvement.

4 Discussion

There is some indication that the three-day forecast error term in the cost function is the main driver of the forecast model improvement. It would be of interest to also investigate this aspect, but that it is beyond the scope of this study.

The parameter uncertainty is specified by expert knowledge as prior values for the first ensemble. It reduces markedly during the estimation process for DETRPEN (Fig. 1), while it increases for the other three parameters. The reasons for this behaviour are twofold. First, the expert uncertainty specification may be too narrow or wide, which then appears as evolving uncertainty. Second, additional experiments (not shown) without stochastic effects indicate that system noise tends to slow down the reduction of parameter uncertainties. Nevertheless, the knowledge about the covariance (Fig. 2b) is new information, and it can potentially guide use of parameter variations as a source of model error in ensemble prediction.

The optimized model was validated in the dependent sample. The EPPES is designed as an online monitoring and parameter estimation tool: by design it is intended to be run as a part of the operational ensemble prediction system, with practically no additional computational cost. Thus, we argue that the primary objective of the EPPES is to perform well in the dependent sample.

The degradation of the 100 hPa geopotential height forecast skill can be attributed to forecast error of mean temperature somewhere between 100 and 500 hPa. However, temperature forecasts both at 500 and 100 hPa verify positively. A possible explanation is the degraded temperature forecast at 200 hPa due to missing O_2 absorption in the radiation scheme. For some unknown reason this feature has a less severe impact in the default model.

The EPPES algorithm passed the critical tests in our experiments. The method was able to improve the forecast skill by tuning low-dimensional static model parameters in high-dimensional Navier–Stokes systems. The proposal distribution for parameter perturbations converged to cover regions where the cost function was improved. The obtained parameter values validated well compared to default values in the already highly tuned ECMWF IFS system. The method effectively integrates out the initial values uncertainty of the state space and the effect of added stochastic noise due to physical tendency perturbations. Furthermore, it is not affected by the problem of collapse of importance-sampling weights. Despite the generally positive result, the optimized model cannot be seriously considered as a "candidate" model for operations. Efforts are needed to formulate a cost function that would lead to such candidate models. Our current thinking is that the target criterion cannot be as selective as the 500 hPa forecast error. Instead, a suitable integral quantity over the entire atmosphere is being searched.

Finally, in the context of NWP, the characteristic parameter distributions (i.e., the distribution parameters $\boldsymbol{\mu}$ and Σ in our case) are not stationary due to, for instance, seasonal and inter-annual variability of the atmosphere. Therefore, one would not expect EPPES, or any other parameter estimation method for that matter, to converge in a strict mathematical sense. This limits the scope of any parameter estimation technique. This holds, in fact, for model tuning even today: models are tuned in limited samples.

5 Summary and conclusions

In this paper, four closure parameters of the ECMWF IFS forecast model at T_L159L62 resolution are estimated using the Ensemble Prediction and Parameter Estimation System (EPPES; Järvinen et al., 2012; Laine et al., 2012; Ollinaho et al., 2013). The estimation procedure is, in short, as follows. The closure parameters are assumed to follow a Gaussian distribution with unknown but static distribution parameters (mean and standard deviation), and the problem is to estimate these distribution parameters instead of the parameters themselves. Standard ensemble predictions are launched, added with initial time parameter variations. Initial state and stochastic physics perturbations are used, just as in the operational ensemble prediction system of ECMWF. The parameter estimation is similar to a sequential application of Bayesian inference, where the likelihood is formulated in terms of three- and ten-day squared forecast error of the 500 hPa geopotential height. The parameter estimation involves 177 ensembles with 50 members from 12 May to 8 August 2011, thus totaling 8850 test forecasts with different parameter combinations.

The parameter mean values increase by about 4–8 % for all four parameters. The posterior distributions indicate noticeable correlation between some parameter pairs. The posterior parameter estimates validate generally positively in a set of 45 forecasts that is a subset of the training set (i.e., a dependent validation sample). In the tropics, the 500 hPa geopotential height mean error, root-mean-squared error, and anomaly correlation coefficient indicate a solid improvement in forecast skill covering almost the entire 10-day forecast range. A scorecard containing a number of scores for various geographical regions, variables, levels, and forecast ranges (in total, 1710 individual scoring elements) also revealed weaknesses. Although the tropical scores were generally improved, even for winds, scores of the 100 hPa geopotential height were markedly degraded. This can be attributed to the selective nature of the likelihood formulation. It is explicit about the three- and ten-day forecast errors at 500 hPa geopotential height, and implicit about errors in mean temperature and humidity in the atmosphere below 500 hPa, plus processes above and below which affect 500 hPa forecast errors. It is not sensitive, however, to height errors (mainly bias) higher up.

Based on the experimentation, the main conclusions are as follows: (i) it is possible to directly tune the predictive skill of a very high dimensional Navier–Stokes system based on ensemble estimation techniques, and (ii) estimation of a small number of model parameters is possible in the presence of stochastic noise due to initial condition and tendency perturbations. The main remaining question is how to formulate the likelihood function such that it leads to a univocal improvement in model performance and its predictive skill. This is our current research topic. Finally, we note that the EPPES computer codes used here are available online at http://helios.fmi.fi/~lainema/eppes.

Acknowledgements. The authors are grateful to Peter Bauer and Anton Beljaars for their support for the research at ECMWF. The research has been funded in part by the Academy of Finland (project numbers 127210, 132808, 133142, and 134999), the Nessling foundation, and by the European Commission's 7th Framework Programme, under grant agreement number 282672, EMBRACE project (www.embrace-project.eu). Support by COST action ES0905 "Basic Concepts for Convection Parameterization in Weather Forecast and Climate Models" is also gratefully acknowledged.

Edited by: P. J. van Leeuwen
Reviewed by: two anonymous referees

References

Bechtold, P., Köhler, M., Jung, T., Leutbecher, M., Rodwell, M., Vitart, F., and Balsamo, G.: Advances in predicting atmospheric variability with the ECMWF model: From synoptic to decadal time-scales, Q. J. R. Meteorol. Soc., 134, 1337–1351, doi:10.1002/qj.289, 2008.

Berner, J., Shutts, G. J., Leutbecher, M., and Palmer, T. N.: A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF ensemble prediction system, J. Atmos. Sci., 66, 603–626, 2009.

Buizza, R., Miller, M., and Palmer, T. N.: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system, Q. J. R. Meteorol. Soc., 125, 2887–2908, 1999.

Buizza, R., Leutbecher, M., and Isaksen, L.: Potential use of an ensemble of analyses in the ECMWF ensemble prediction system, Q. J. R. Meteorol. Soc., 134, 2051–2066, 2008.

de Rooy, W. C., Bechtold, P., Fröhlich, K., Hohenegger, C., Jonker, H., Mironov, S., Teixeira, J., and Yano, J.-I.: Entrainment and detrainment in cumulus convection: an overview, Q. J. R. Meteorol. Soc., 139, 1–19, doi:10.1002/qj.1959, 2013.

Isaksen, L., Bonavita, M., Buizza, R., Fisher, M., Haseler, J., Leutbecher, M., and Raynaud, L.: Ensemble of Data Assimilations at ECMWF, ECMWF Tech. Mem., 636, 46 pp., 2010.

Järvinen, H., Laine, M., Solonen, A., and Haario, H.: Ensemble prediction and parameter estimation system: the concept, Q. J. R. Meteorol. Soc., 138, 281–288, doi:10.1002/qj.923, 2012.

Kivman, G. A.: Sequential parameter estimation for stochastic systems, Nonlin. Processes Geophys., 10, 253–259, doi:10.5194/npg-10-253-2003, 2003.

Laine, M., Solonen, A., Haario, H., and Järvinen, H.: Ensemble prediction and parameter estimation system: the method, Q. J. R. Meteorol. Soc., 138, 289–297, doi:10.1002/qj.922, 2012.

Lorenz, E. N.: Predictability: A problem partly solved. Proceedings of the Seminar on Predictability, Vol. I, ECMWF, Reading, UK, 1–18, available at: www.ecmwf.int/publications, 1995.

Ollinaho, P., Järvinen, H., Laine, M., Solonen, A., and Haario, H.: NWP model forecast skill optimization via closure parameter variations, Q. J. R. Meteorol. Soc., 139, 1520–1532, doi:10.1002/qj.2044, 2013.

Palmer, T. N., Buizza, R., Doblas-Reyes, F., Jung, T., Leutbecher, M., Shutts, G. J., Steinheimer, M., and Weisheimer, A.: Stochastic parameterization and model uncertainty, ECMWF Tech. Memo., 598, 42 pp., 2009.

Roeckner, E., Bäuml, G., Bonaventura, L., Brokopf, R., Esch, M., Giorgetta, M., Hagemann, S., Kirchner, I., Kornblueh, L., Manzini, E., Rhodin, A., Schlese, U., Schulzweida, U., and Tompkins, A.: The atmospheric general circulation model ECHAM5, Part I Model Description, Tech. Rep. No. 349, Max-Planck-Institut fur Meteorologie, 2003.

Rougier, J.: "Intractable and unsolved": some thoughts on statistical data assimilation with uncertain static parameters, Phil. Trans. R. Soc. A, 371, 20120297, 10.1098/rsta.2012.0297, 2013.

Simmons, A. J. and Hollingsworth, A.: Some aspects of the improvement in skill of numerical weather prediction, Q. J. R. Meteorol. Soc., 128, 647–677, doi:10.1256/003590002321042135, 2002.

Tiedtke, M.: A comprehensive mass flux scheme for cumulus parameterization in large-scale models, Mon. Weather Rev., 117, 1779–1800, 1989.

van Leeuwen, P. J.: A Variance-Minimizing Filter for Large-Scale Applications, Mon. Weather Rev., 131, 2071–2084, doi:10.1175/1520-0493(2003)131<2071:AVFFLA>2.0.CO;2, 2003.

Wilks, D. S.: Effects of stochastic parametrizations in the Lorenz '96 system, Q. J. R. Meteorol. Soc., 131, 389–407, 2005.