Lauri Tuppi[*1]
Pirkka Ollinaho[2], Madeleine Ekblom[1],
Vladimir Shemyakin[3], Heikki Järvinen[1]

[1]Institute for Atmospheric and Earth System Research / Physics, University of Helsinki, Helsinki, Finland
[2]Finnish Meteorological Institute, Helsinki, Finland
[3]Lappeenranta University of Technology, Lappeenranta, Finland

*Correspondence to lauri.tuppi@helsinki.fi

NECESSARY CONDITIONS FOR ALGORITHMIC TUNING OF WEATHER MODELS USING OPENIFS AS AN EXAMPLE

1. INTRODUCTION

- Manual tuning of NWP models is laborious and lacks transparency (Mauritsen et al 2012, Hourdin et al. 2017) → usage of algorithmic methods
- Little studies done about how to use algorithmic tuning methods in the best way
- We study parameter convergence in “convergence tests”: unperturbed control model used as reference
- We considered following aspects:
  - selecting level of realism
  - selecting optimisation target
  - maximising computational efficiency
  - reproducibility of the results
  - trustability of algorithmic tuning
  - potential pitfalls
- Highlights of Tuppi et al. 2020 are presented

2. EXPERIMENT SET-UPS AND TOOLS

- Closure parameters of the convection scheme
- OpenEPS ensemble prediction workflow manager (Ollinaho et al. in prep.)
- Tuning algorithms embedded in OpenEPS: EPPES (Järvinen et al. 2012, Laine et al. 2012) and DE (Shemyakin and Haario 2018)
- Optimisation targets: RMS error of Z850 ΔZ and moist total energy norm ΔE_m (e.g. Ehrendorfer et al. 1999):

$$\Delta Z = \sqrt{\frac{1}{D}\int_D (Z_{850} - Z'_{850})^2 dD}$$

$$\Delta E_m = \frac{1}{2}\int_\eta \int_D [u'^2 + v'^2 + c_q \frac{L^2}{c_p T_r} q'^2] dD \frac{\delta p_r}{\delta \eta} d\eta + \frac{1}{2}\int_D [R\frac{T_r}{p_r} \ln p_s'^2] dD,$$

- Fair CRPS (Leutbecher 2018) for measuring the convergence

$$\mathrm{fCRPS}_1 = \frac{1}{M}\sum_{j=1}^{M} |\theta'_{j,n} - \theta_{d,n}| \qquad \mathrm{fCRPS}_2 = \frac{1}{2M(M-1)}\sum_{i=1}^{M}\sum_{k=1}^{M} |\theta'_{j,n} - \theta'_{k,n}|,$$

- Convergence tests with different levels of realism:

	Number of parameters	Different initial conditions	Stochastic physics (SPPT)
Level 0 (L0)	2	No	No
Level 1 (L1)	2	Yes	No
Level 2 (L2)	2	Yes	Yes
Level 3 (L3)	5	Yes	Yes

3. RESULTS

- Moderate level of realism, this increases sample diversity (Fig. 1, S1)
- Cost function: more comprehensive, the better
- Computational efficiency: short forecasts of 24h best in this example (Fig. 2, S3), increasing ensemble size not accelerating convergence (Fig. 3, S4)
- Convergence with 24h forecasts the most reproducible (S5)
- Algorithmic tuning can be trusted but careless tuning leads to bad results; take it as expert-guided

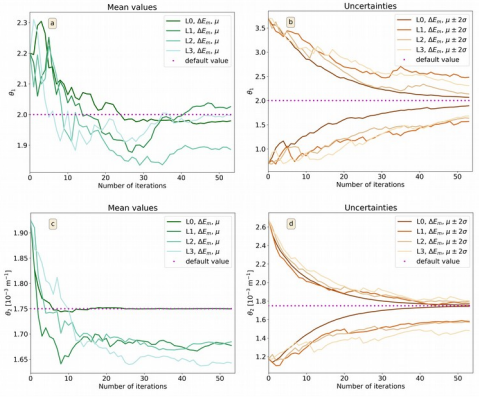
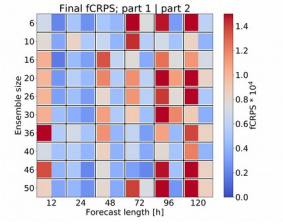


Figure 1. Convergence with different levels of realism for parameter 1 in (a, b) and for parameter 2 in (c, d). Ensemble size is 50 members and forecast range 48h.



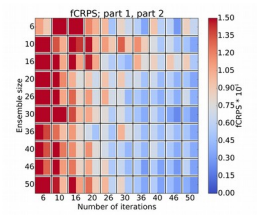


Figure 2. L2 convergence test shown for parameter θ_2 with different forecast ranges and ensemble sizes. Convergence measured with fCRPS_1 and fCRPS_2. Perfect convergence is zero.

Figure 3. Evolution of convergence of θ_2 with different ensemble sizes. Forecast range is 24 hours.

4. RECIPE FOR SUCCESFUL TUNING

- Moderate level of realism (initial condition perturbations + possibly model stochastic perturbations)
- Comprehensive measure used as cost function (ΔE_m in our case)
- A relatively short forecast range (24 hours in our case)
- A relatively small ensemble size (20 in our case)
- Tested with 8 parameters. All parameters converged toward the default values (Fig. S7)

5. REFERENCES

Ehrendorfer et al. 1999, doi:10.1175/1520-0469(1999)056<1627:SVPGIA>2.0.CO;2
Hourdin et al. 2017, doi:10.1175/BAMS-D-15-00135
Järvinen et al. 2011, doi:10.1002/qj.923
Laine et al. 2012, doi:10.1002/qj.922
Leutbecher 2018, doi: 10.1002/qj.3387
Mauritsen et al. 2012, doi:10.1029/2012MS000154
Ollinaho et al. 2020, A year of ensemble initial conditions, in prep.
Shemyakin and Haario 2018, doi:10.1007/s11071-018-4239-5
Tuppi et al. 2020, Necessary conditions for algorithmic tuning of weather models using OpenIFS as an example, submitted to GMD

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI
MATEMAATTIS-LUONNONTIETEELLINEN TIEDEKUNTA
MATEMATISK-NATURVETENSKAPLIGA FAKULTETEN
FACULTY OF SCIENCE

Supplementary material for the poster

	Number of parameters	**Different initial conditions**	**Stochastic physics (SPPT)**
Level 0 (L0)	2	No	No
Level 1 (L1)	2	Yes	No
Level 2 (L2)	2	Yes	Yes
Level 3 (L3)	5	Yes	Yes

Table 1. Different levels of realism used in the convergence test.

Root mean squared error of 850 hPa geopotential:

$$\Delta Z = \sqrt{\frac{1}{D}\int_D (Z_{850} - Z'_{850})^2 dD} \qquad (1)$$

Z_{850} = 850 hPa geopotential at a grid point in the control forecast
Z'_{850} = 850 hPa geopotential at a grid point in a perturbed forecast
D = horizontal domain

Moist total energy norm (e.g. Ehrendorfer et al. 1999):

$$\Delta E_m = \frac{1}{2}\int_\eta \int_D [u'^2 + v'^2 + c_q \frac{L^2}{c_p T_r} q'^2] dD \frac{\delta p_r}{\delta \eta} d\eta + \frac{1}{2}\int_D [R\frac{T_r}{p_r} \ln p_s'^2] dD, \qquad (2)$$

u' = difference of u wind between the control and perturbed forecast at a grid point
v' = difference of v wind between the control and perturbed forecast at a grid point
q' = difference of specific humidity between the control and perturbed forecast at a grid point
p'_s = difference of surface pressure between the control and perturbed forecast at a grid point
c_q = scaling constant for the moisture term, (we use c_q=1)
L = vaporisation energy of water
c_p = specific heat constant of air at constant pressure
T_r = reference temperature, (we use T_r = 280 K)
p_r = reference pressure, (we use 1000 hPa)
$\delta p_r/\delta\eta$ = difference of pressure between two model levels
D = horizontal domain
η = vertical domain

Bias part of the fair continuous ranked probability score (see Leutbecher 2018):

$$\mathrm{fCRPS}_1 = \frac{1}{M}\sum_{j=1}^{M}|\theta'_{j,n} - \theta_{d,n}| \qquad (3)$$

Spread part of the fair continuous ranked probability score (see Leutbecher 2018):

$$\mathrm{fCRPS}_2 = \frac{1}{2M(M-1)}\sum_{i=1}^{M}\sum_{k=1}^{M}|\theta'_{j,n} - \theta'_{k,n}|, \qquad (4)$$

M = ensemble size
$\theta'_{j,n}$ = perturbed parameter value of ensemble member j and parameter n
$\theta_{d,n}$ = default value of parameter n
$\theta'_{k,n}$ = perturbed parameter value of ensemble member k and parameter n

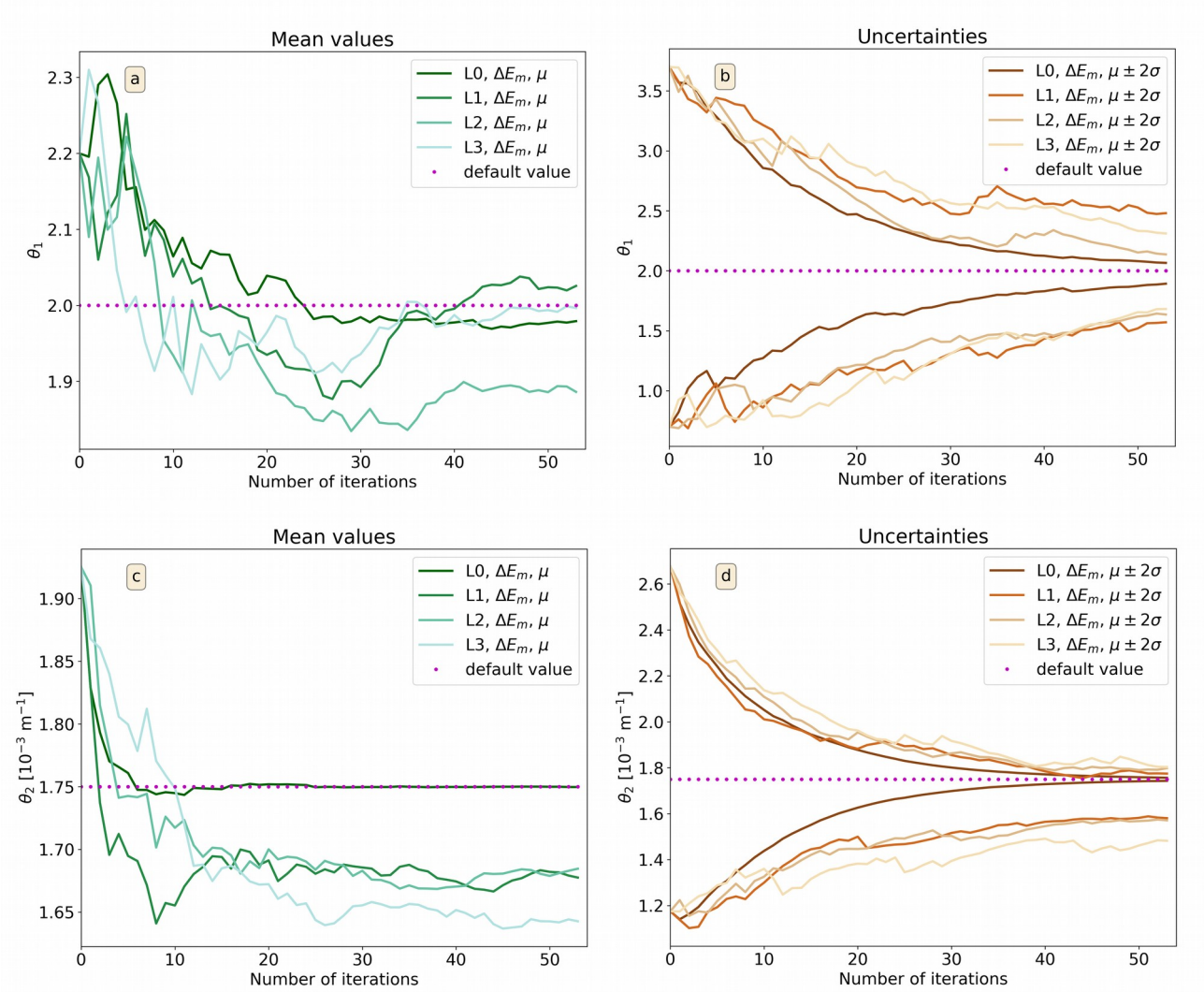


Figure S1: Comparison of convergence tests at different levels of realism. Panels (a) and (b) show the evolution of distribution mean value (μ) and the mean value ±2 standard deviations uncertainty (μ ± 2σ) for parameter θ_1, and (c) and (d) show the same as (a) and (b) but for parameter θ_2. The purple dots show the parameter default values. The x-axes show running number of iterations, i.e., how many ensemble forecasts that have been used. Moist total energy norm (ΔE_m) is used as the cost function, and the levels of realism are summarised in Table 1. EPPES is used as the optimiser, the ensemble size is 50 members and the forecast range 48 hours. (Figure 1 of Tuppi et al. 2020).

Interpretation. We think that L0 test do not reveal the true performance in fully-realistic model tuning, which uses analyses or observations as reference data. In fully-realistic tuning the truth is unknown so using L1 is safe enough simple option. Initial condition perturbations make the initial condition sample more diverse so we expect those perturbations to enhance parameter convergence in fully realistic tuning. Higher levels of realism did not affect the convergence test results much so we expect them to have relatively small impact also in fully-realistic tuning.

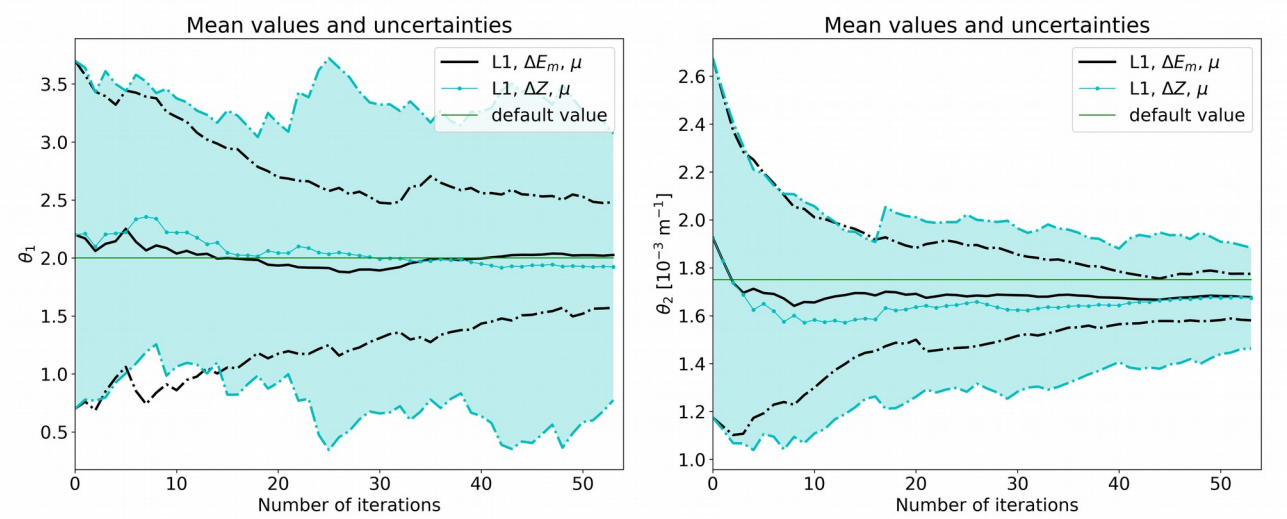


Figure S2: Convergence tests with different cost functions. Convergence of θ_1 on the left and θ_2 on the right. The x-axes show running number of iterations. Solid black lines show the evolution of distribution mean values (μ) and black dash-dotted lines the mean values ±2 standard deviations when ΔE_m is used as cost function. Cyan dotted lines and shading in the background show the same for ΔZ. Default value shows the fixed parameter value used in the control model. Both convergence tests are L1 tests with 50 ensemble members and 48 hour forecasts. EPPES is used as optimiser. (Figure 2 of Tuppi et al. 2020).

Interpretation. More comprehencive cost function (ΔE_m) leads to faster and more reliable convergence. Single-variable and single-level cost functions such as ΔZ constrain the quite poorly so the signal caused by parameter perturbations may be too weak to be detected. As an example, perturbing entrainment of shallow convection is almost not at all visible in 850 hPa geopotential height in 48 hour forecasts but it is very visible in specific humidity field.

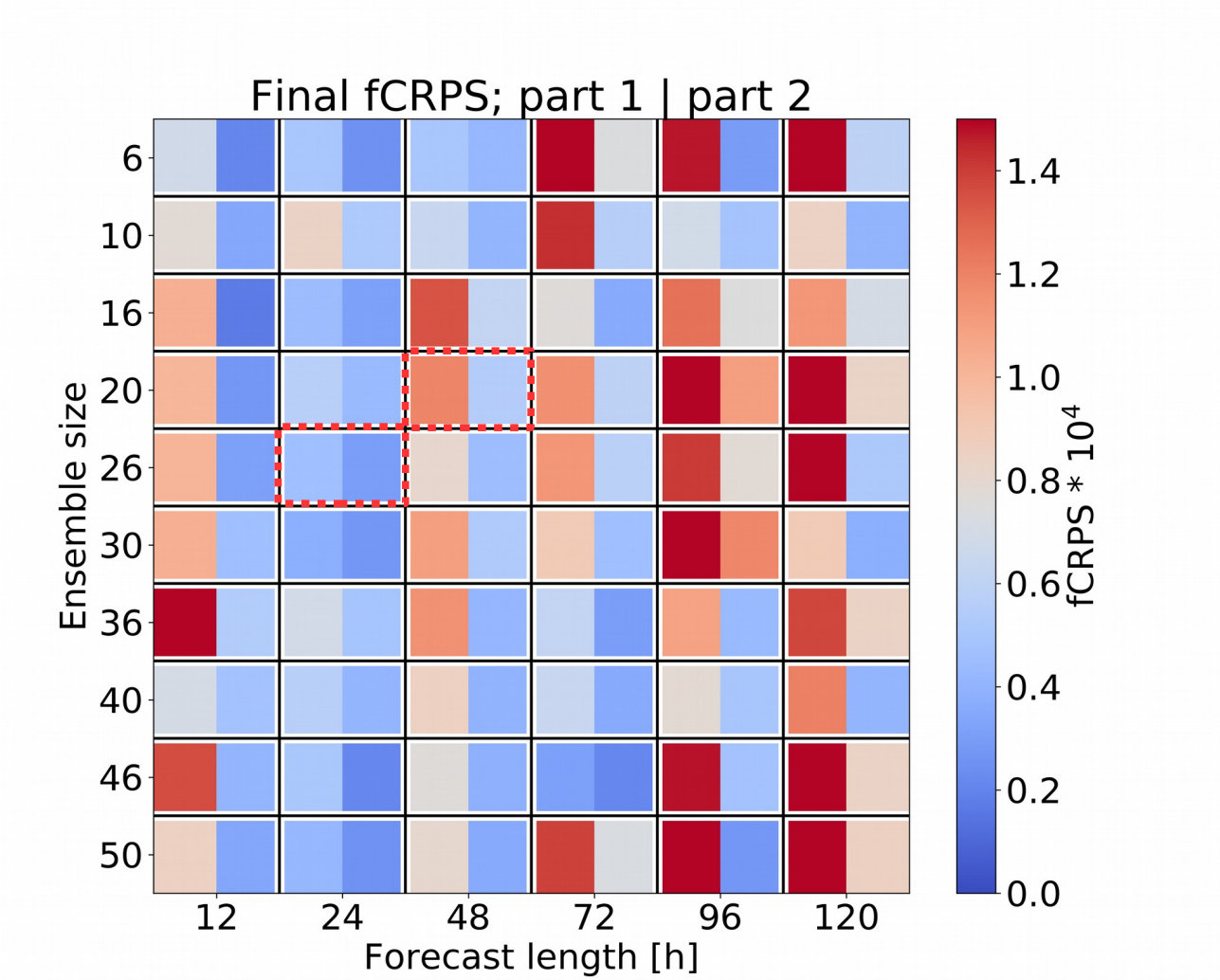


Figure S3: Components of fair CRPS from the final iteration of the convergence tests with various forecast ranges and ensemble sizes. In this example, the optimisation algorithm is EPPES and the parameter is θ_2. The left-hand side of each block repre- sents the average distance of the parameter values from the default value (equation 3), and the right-hand side represents the spread of the parameter value distribution (equation 4). Low values and blue colours of both sides of the blocks indicate good convergence. (Figure 3 of Tuppi et al. 2020).

Interpretation. Low values of both components of fCRPS denote good convergence. Blue colour of part 1 means that the parameter mean value is close to the default value, and blue colour of part 2 means that the parameter distribution is very narrow. Perfect convergence would mean zero scores as then the parameter distribution would be shrunken to a dot exactly on the default value. 24 hour forecast forecast range is the most optimal forecast range for θ_2. For the other parameter θ_1 12h, 24h and 48h ranges are equally optimal whereas longer ranges are clearly suboptimal (not shown). The reason why part 1 (the bias part) is more often red than part 2 (the spread part) is explained in figure S6 and the interpretation therein.

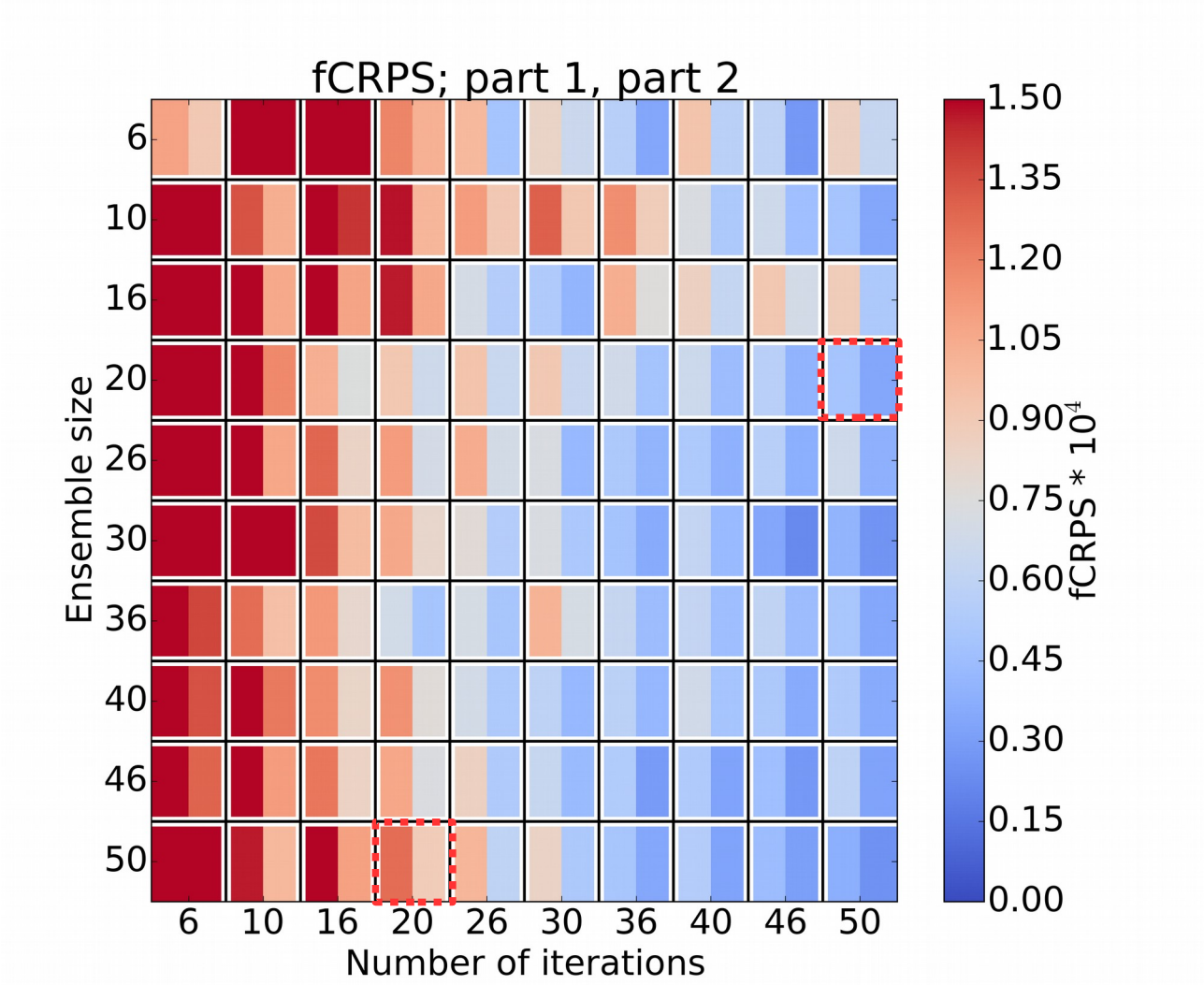


Figure S4: Evolution of fCRPS of parameter θ_2 in convergence tests with L2, ΔE_m , EPPES, 24 hour forecasts and various ensemble sizes. The interpretation of the blocks is the same as in Figure S3. The number of iterations indicates how many iterations of the algorithm have been done, or in other words how many ensemble forecasts have been run. Components of fCRPS have been calculated using equations (3) and (4). (Figure 4 of Tuppi et al. 2020).

Interpretation. Even though the ensemble size varies substantially, the rate of convergence stays roughly constant. Therefore, using large ensemble size can be seen as unmeaningful burning of computer resources. Example: convergence tests with 20 member and 50 iterations, and 50 members and 20 iterations have both used 1000 forecasts with parameter perturbations (highlighted in Figure S4). However, the former option leads to much better convergence. Instead using very small ensembles was observed to make the convergence tests occasionally unstable so we recommend to use moderate ensemble size of some 20 members.

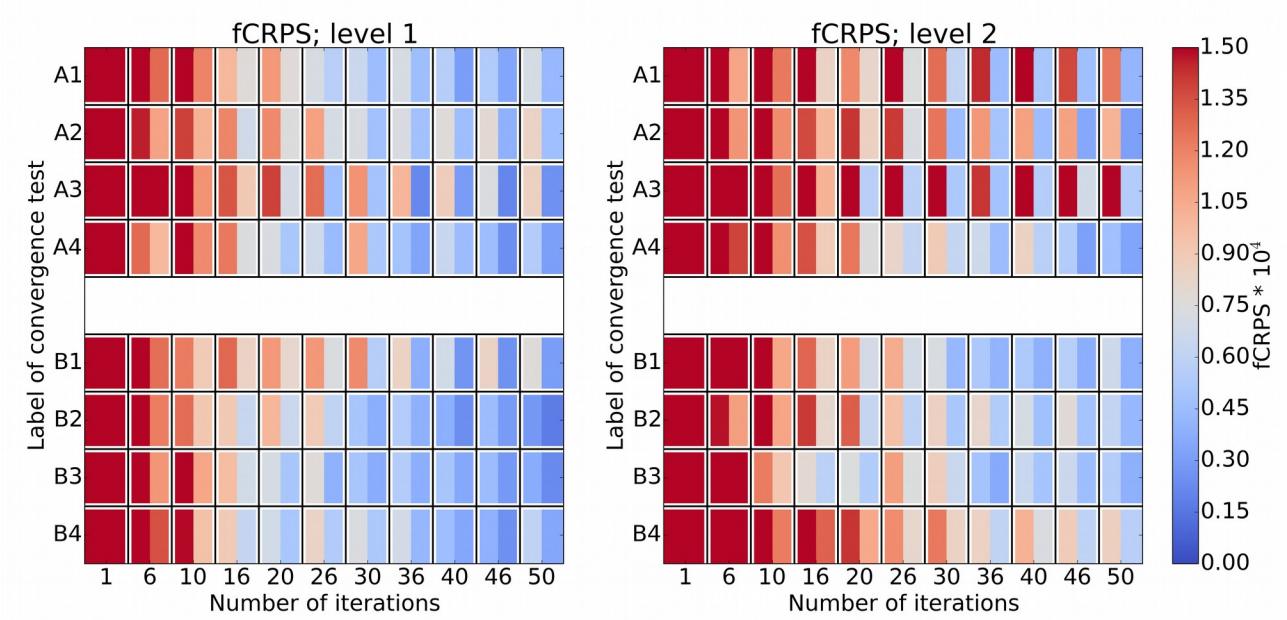


Figure S5: Evolution of θ_2 in repeated convergence tests with two selected forecast range – ensemble size combinations highlighted in Figure S3. The level of realism is L1 on the left and L2 on the right. Tests A1 to A4 have been run with forecast range of 48 hours and ensemble size of 20 members, and tests B1 to B4 with 24 hours and 26 members. EPPES was used as an optimiser in these examples. Components of fCRPS have been calculated using equations (3) and (4). (Figure 5 of Tuppi et al. 2020).

Interpretation. Both optimal and suboptimal combinations lead to reproducible convergence with L1 as is shown in the left-hand side panel. Those convergence tests produce relatively similar results every time. Instead, in L2 tests in the right-hand side panel only the optimal combination labelled with B1 to B4 yields reproducible convergence. This gives further support that 24 hours is optimal forecast range at least for parameter θ_2. Results with θ_1 were less conclusive (not shown).

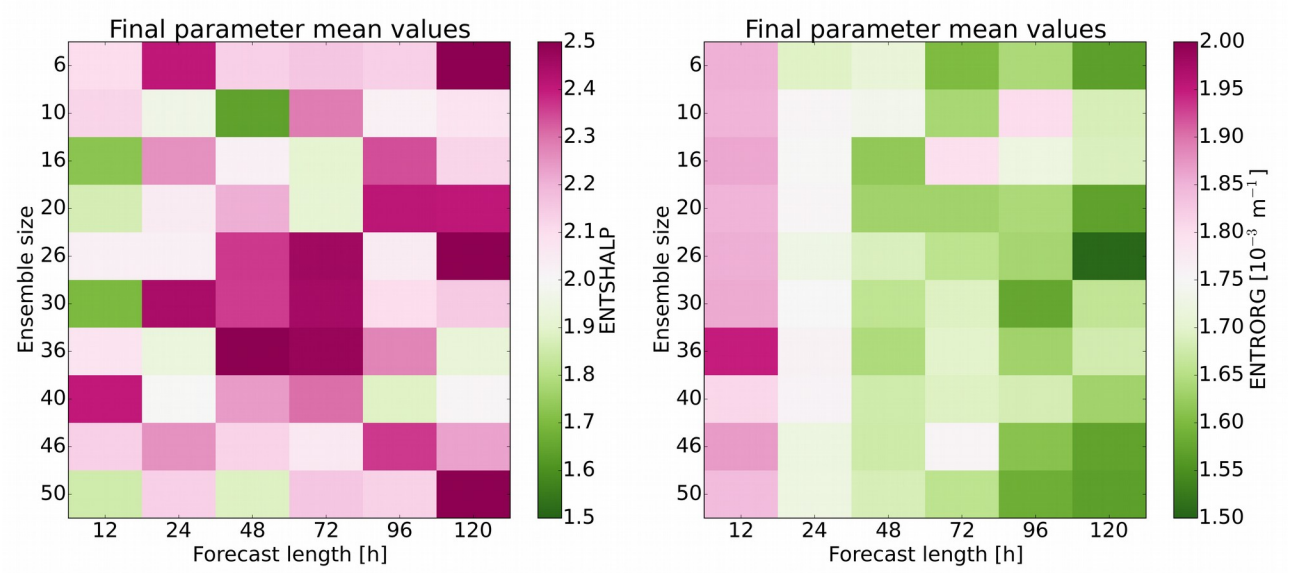


Figure S6: Mean values of the parameter distributions proposed by EPPES at the end of the convergence tests. Mean values of θ_1 are on the left and mean values of θ_2 on the right. Purple (green) colour means that the final mean values are larger (smaller) than the default value. (Figure 6 of Tuppi et al. 2020).

Interpretation. Optimal parameter values seem to depend on the forecast range used. We tested some of those parameter values proposed by EPPES, and run regular ensemble forecasts, and then compared the forecasts to the control forecasts with ΔE_m. Indeed the cost function values were lower than when default parameter values were used. Therefore this phenomenon is not caused by the tuning infrastructure but it is actually a sign that model closure parameters and properties of ensemble forecasts are connected. Most likely perturbing closure parameters affects how much the ensemble forecasts can generate spread here so that the optimal parameter values lead to slightly smaller spread than the default values. This is not a problem for algorithmic ensemble based tuning but a feature, which should be kept in mind.

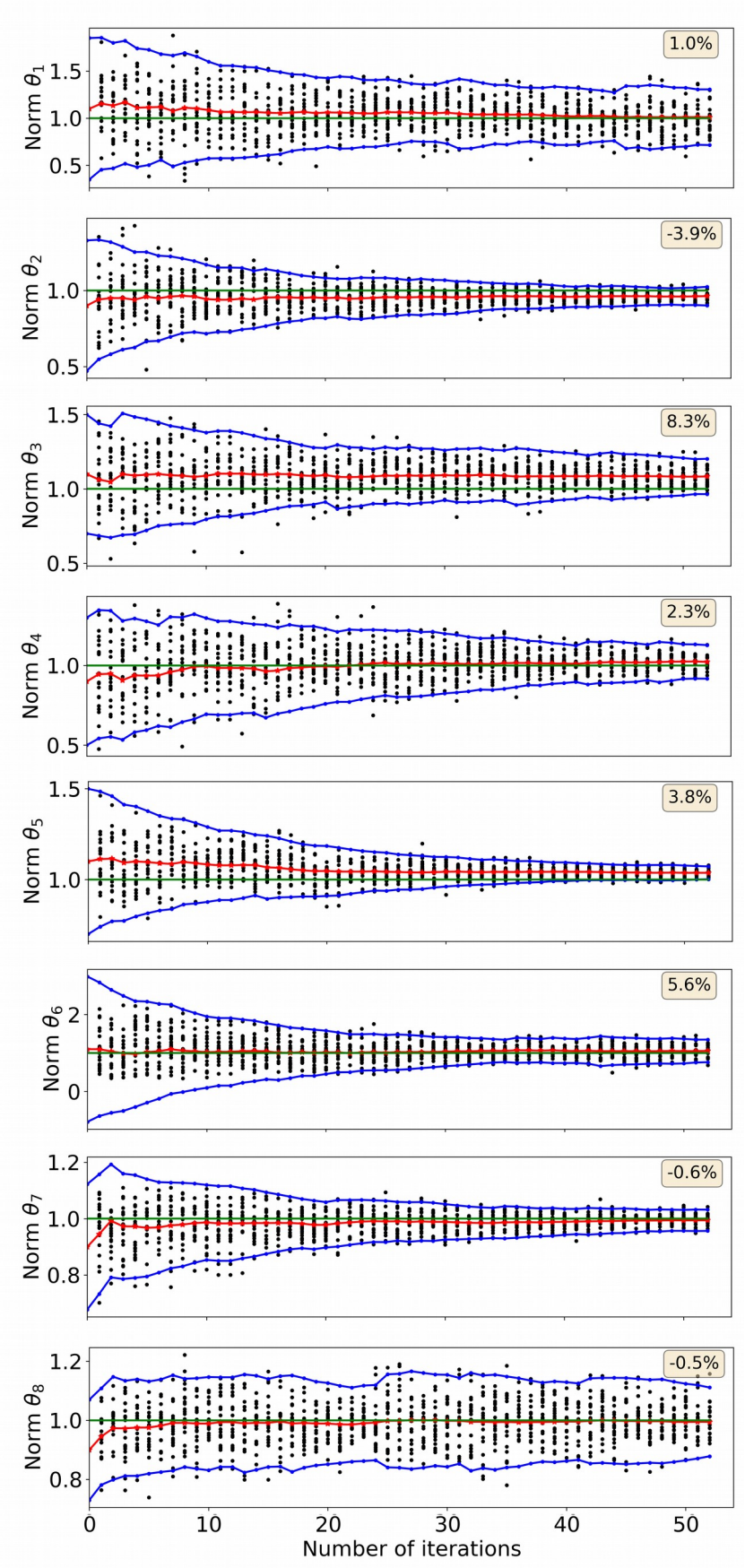


Figure S7: Progress of the convergence in the eight-parameter test. The parameter values and uncertainties have been normalised with their default values. Black dots show sampled

parameter values, red line with stars shows parameter mean value, blue lines with dots show mean value ±2 standard deviations and the green line shows the default parameter value that is 1.0 due to the normalisation. The text boxes indicate the remaining parameter off-set, which is the relative distance between the final parameter mean value and the default value. Initial parameter off-set is randomly plus or minus 10 %. Forecast range is 24 hours and ensemble size 20 members in this L1 convergence test. ΔE_m is used as the cost function and EPPES as the optimiser. (Figure 7 of Tuppi et al. 2020).

Interpretation. All eight parameters converge satisfactorily. Added dimensionality does not seem to hinder EPPES to find the parameter default values. Some off-set is often left but the off-set is not statistically significant as the default value stays within the uncertainty. Successful convergence test with eight parameters raises the probability that fully-realistic tuning of entire weather model with some 20 parameters at once is possible.

References:

Ehrendorfer M. et al. Singular-vector perturbation growth in a primitive equation model with moist physics. Journal of the Atmospheric Sciences, 56(11):1627–1648, 1999, doi:10.1175/1520-0469(1999)056<1627:SVPGIA>2.0.CO;2

Leutbecher M., Ensemble size: How suboptimal is less than infinity? Quarterly Journal of the Royal Meteorological Society, 0(0), 2018. 2018, doi: 10.1002/qj.3387

Tuppi et al. 2020, Necessary conditions for algorithmic tuning of weather models using OpenIFS as an example, submitted to Geoscientific Model Development