LACE stay, ZAMG

24. 6. – 19. 7. 2019.

Adding lagged deterministic forecasts to a convection-permitting EPS

Plus: V-matrix recalculation and solving problems with 903 and C-LAEF

Author:

Endi Keresturi

endi.keresturi@cirus.dhz.hr

Supervisors:

Christoph Wittmann
Clemens Wastl

Preface

This report consists of 4 main parts:

1) Adding lagged deterministic forecasts to a convection-permitting EPS
2) Resolving problems in C-LAEF
3) Recalculation of the V-matrix
4) 903 bug or a feature?

We now proceed with part 1)

Introduction

At ZAMG, Convection-permitting - Limited Area Ensemble Forecasting system (C-LAEF) is being developed (Wastl *et al.*, 2019; Keresturi *et al.*, 2019). C-LAEF is based on AROME model and it is currently run in a preoperational mode (since June 2019) with a horizontal grid spacing of 2.5 km and 90 vertical levels. It comprises 16 members (plus control) using the first 16 out of a total of 51 members of ECMWF-EPS for the boundary conditions/uncertainties. For representing model error, a newly developed hybrid stochastic perturbation scheme is applied in C-LAEF (Wastl *et al.*, 2019). In this scheme tendencies perturbations in shallow convection, radiation and microphysics are combined with parameter perturbations in the turbulence scheme. For representing initial condition uncertainties, 3D-Var EDA with ensemble *Jk* method (Keresturi *et al*., 2019) and observation perturbations in CANARI are used. Assimilation cycles are performed every 6 h with 48 h forecasts issued at 00 and 06 UTC.

Many studies in the past have dealt with the idea of extending an Ensemble Prediction System (EPS) or even a deterministic system by cheap techniques like lagging (Hoffmann and Kalnay, 1983; Bouallegue *et al.*, 2013), neighbourhood approaches (Theis *et al.*, 2005; Bouallegue *et al.*, 2013) or combination of many different deterministic models/EPSs (Bowler et al., 2007, Buizza 2014). Regarding the Limited Area EPS (LAMEPS), Raynaud and Bouttier (2017) focused on extending LAMEPS by adding lagged members from the same EPS, while Mittermaier (2007) explored possibility of creating an LAMEPS from lagged runs of a deterministic model. All before mentioned studies reported benefits when using before mentioned techniques.

In this study, we explore a different possibility – extending an LAMEPS with lagged deterministic forecasts. This unique setup is possible because, at ZAMG, operational deterministic AROME system is configured the same (except for 3 h cycling) as C-LAEF (C-LAEF control to be precise). AROME forecast range is 60 h which enables us to use up to 4 older AROME runs and combine them with C-LAEF to create a new 21-member ensemble and keep 48 h forecast range. Because of this, we expect to have an interchangeable 21 member EPS where all members are equally likely.

Special weighting strategy will not be applied and all members will be weighted equally as previous studies have already shown its negligible impact on time-lagged ensembles (Bouallegue *et al.*, 2013; Raynaud *et al.*, 2015). For our case this is even more valid as we are only adding 4 additional members.

Model setup and experiments

Our goal is to assess the added value of including 4 lagged AROME members to the C-LAEF ensemble. For this reason, we define two experiments: a) REF – represents C-LAEF raw ensemble (17 members) and b) LAG – C-LAEF plus 4 lagged AROME runs (-3, -6, -9 and -12 h). Both AROME and C-LAEF are configured as described in the introduction. Lagged members and C-LAEF control are coupled to ECMWF HRES. Domain of integration is shown in Fig. 1.

Verification

C-LAEF 00 UTC runs (hourly output) were archived for the period 19. 5. – 30. 6. 2019. with two missing days (26. 5. and 2. 6.). This gives us the total of 41 days for which to perform a verification. Forecasts for variables were archived – precipitation (RR), wind speed (WS), wind gusts (WG) and surface global short-wave radiation (GR; "*SURFRAYT SOLA DE*"). Verification was performed against 320 automatic surface stations within Austria by matching nearest model grid point to the observation location. For precipitation, INCA analyses upscaled by bilinear interpolation to the model grid were used instead. To determine if the difference in scores between the experiments is statistically significant, the bootstrap technique using 1,500 re-samples was applied.

In order to correctly evaluate an EPS, one must consider different aspects of it (e. g. attributes), such as accuracy, reliability, resolution, discrimination, etc. (Murphy, 1993). Given these sources of complexity, no single score can assess all attributes of probabilistic forecasts at once. For this reason, various scores are used to assess different aspects of forecasts. These include (a) the root-mean-square error (RMSE) of ensemble mean and Brier score for assessing ensemble mean accuracy, (b) the continuous rank probability score (CRPS) for assessing overall EPS skill, (c) ensemble RMSE/spread relation and outlier statistics for assessing reliability, (d) relative operational characteristics (ROC) for assessing discrimination and (e) Brier score decomposition for assessing resolution (and reliability).

As it is well known and understood, precipitation should not be verified in this "traditional" way (Ebert, 2008; Mittermaier, 2014). For this reason, precipitation is also verified by using a form of neighbourhood approach - Fraction Skill Score (FSS; Roberts and Lean, 2008). FSS is computed following the ensemble formulation proposed by Duc *et al*. (2013).

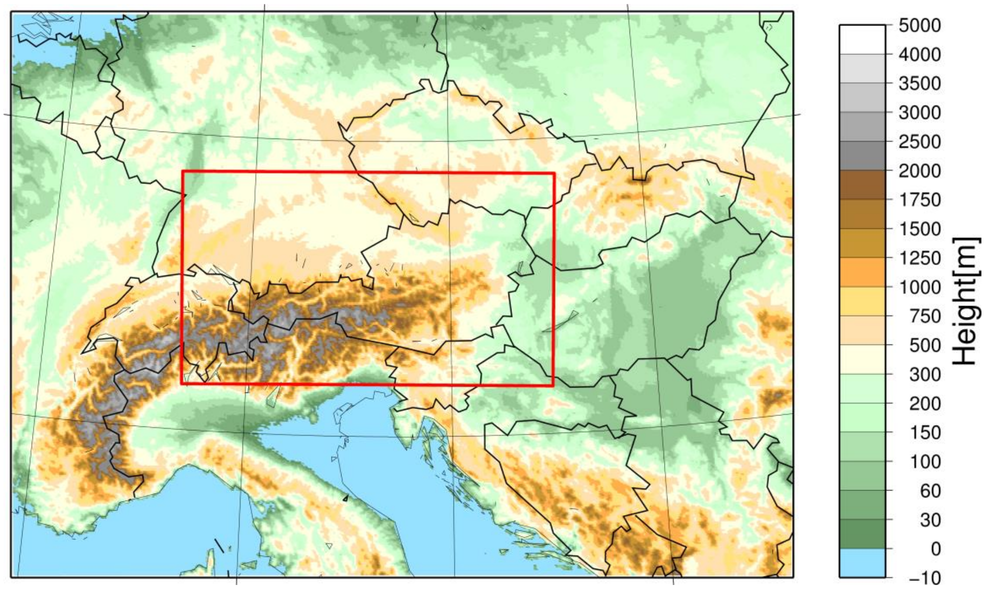

Figure 1. C-LAEF computational domain and orography. Area outlined marks the INCA domain.

Decomposition of CRPS to a term related to spread and a term related to mean absolute error (Leutbecher, 2018), and decomposition of RMSE to BIAS, BIAS of standard deviation and dispersion (Horvath e*t al*., 2012) will help us take a deeper insight into the reasons behind the obtained results.

Results and discussion

Verification for surface global short-wave radiation gave neutral results for all scores and will not be discussed further. Also, traditional verification for precipitation yielded similarly neutral results for most of the scores and will not be discussed further, except for few cases that will be mentioned. Precipitation results will be discussed in more detail in terms of FSS. As we are adding only 4 extra members, we cannot expect too much difference between the experiments.

We will now present results obtained by averaging over the entire verification period of 41 days.

a) RMSE of ensemble mean and spread

Fig. 2 shows RMSE and spread of REF and LAG for WS and WD. As we can see, there is hardly an impact on accuracy of the ensemble mean for WS and WG, while spread is significantly (on 95 % level) increased in both cases implying better reliability of LAG. Decomposition of RMSE revealed that BIAS of WS (WD) is slightly reduced (increased)

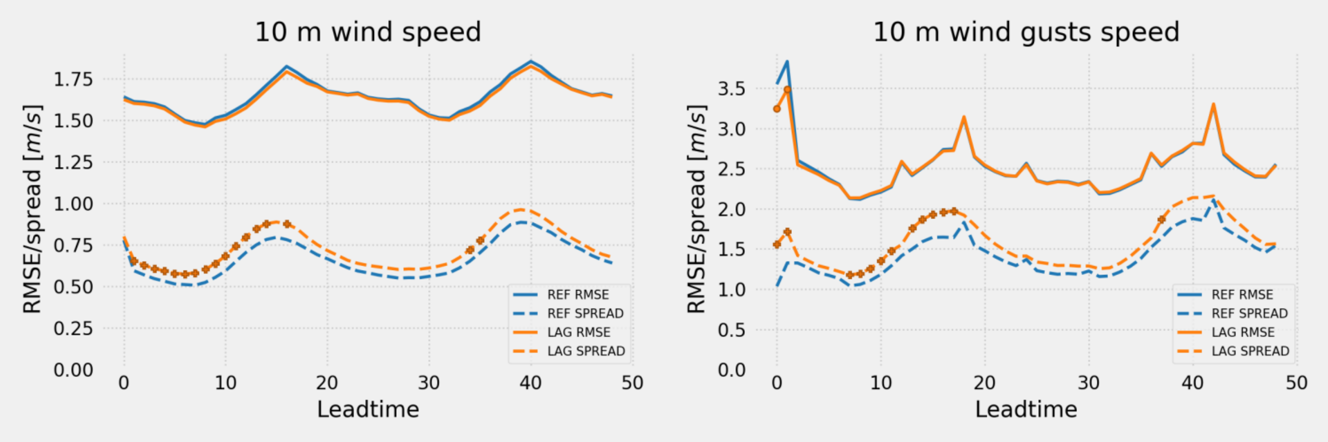

Figure 2. RMSE of the ensemble mean and ensemble spread for WS (left) and WG (right). Forecast ranges with statistically significant differences are marked with bullets.

in LAG (not shown). This increase in BIAS is countered by lower dispersion error in LAG (not shown).

Looking more thoroughly at WD plots, one can observe strange behaviour (peaks) of RMSE curve at +0 and +1 h, but also at +6, +12 … at 6 h intervals. Those were identified as **problem 1** and **2** in the pre-operational C-LAEF forecasts and will be discussed later.

b) *CRPS*

Fig. 3 shows CRPS and its decomposition for WS and WD. We see a slight reduction of CRPS in LAG for both variables, although not significant on 95 % level. Decomposition shows us that this reduction of CRPS comes from significant increase in ensemble spread in LAG.

c) *Outlier statistics*

From Fig. 4, it is very clear that number of outliers is reduced in LAG meaning reliability is increased. This is consistent with results from *a).*

d) *ROC*

Fig. 5 shows ROC curve and ROC area for thresholds of 3 and 5 m/s (10 and 15 m/s) for WS (WG). For all of 4 combination, discrimination is slightly improved in LAG. Here, we have also observed improved discrimination in LAG for precipitation on all thresholds tested (1, 3, 5 and 10 mm; not shown).

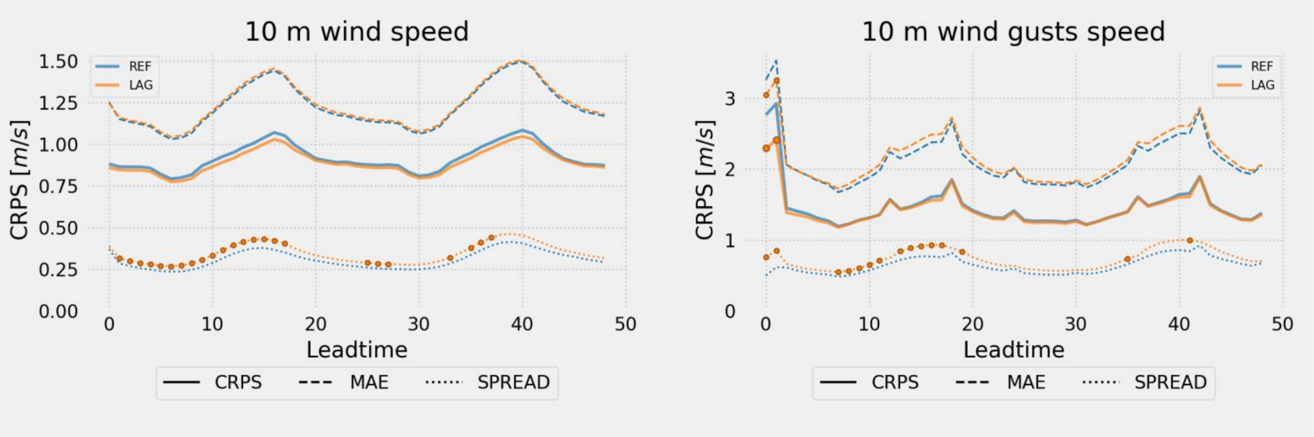

Figure 3. Decomposition of CRPS for WS (left) and WG (right). Forecast ranges with statistically significant differences are marked with bullets.

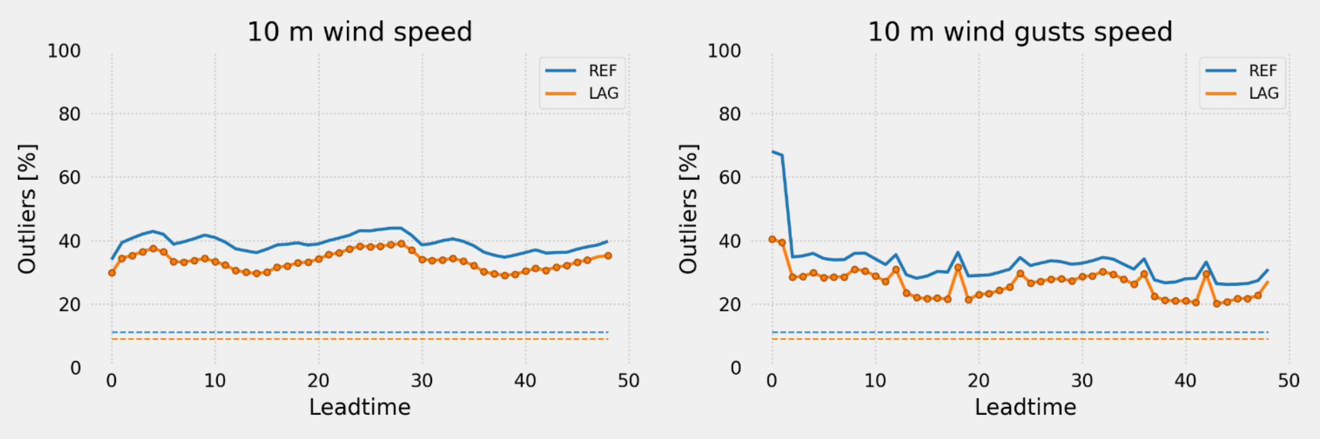

Figure 4. Percentage of outliers for WS (left) and WG (right). Forecast ranges with statistically significant differences are marked with bullets. Dashed horizontal lines denote ideal values.

e) Brier score

Decomposition of Brier score confirms previous conclusions that LAG is more reliable ensemble, while resolution is similar for both experiments (not shown).

f) Precipitation

The left panel of Fig. 6 shows the median of the Skill Score of FSS of LAG to FSS of REF illustrated as a matrix of colours (red means LAG is better than REF and blue the opposite, white is for no difference) for different thresholds, scales and forecast ranges.

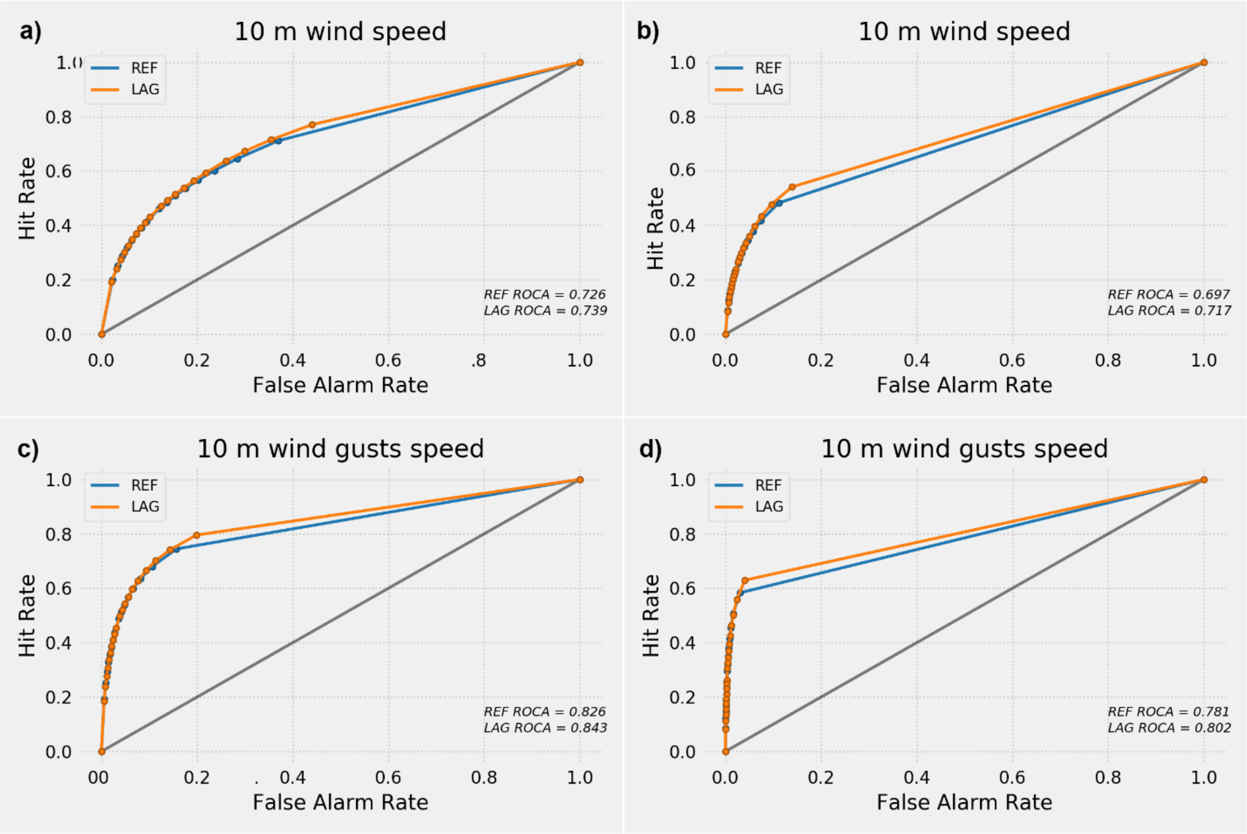

Figure 5. ROC curve and ROC area for: (a) WS (3 m/s), (b) WS (5 m/s), (c) WG (10 m/s) and (d) WG (15 m/s).

The right panel we can see the percentage of times (days) FSS of LAG is higher than the FSS of REF. FSS is visibly improved in LAG for the first 5 h of integration mostly for scales higher than 25 km and thresholds up to 2 mm. Small improvements are still visible up to 20 h of integration, after which results become mostly neutral.

It is interesting to ask wheatear the observed differences between the experiments are only due to the larger ensemble size of LAG? To answer this question, verification was performed again, but this time 4 C-LAEF members were removed in LAG, so that both experiments have 17 members. All of the results and conclusions remained the same (not shown) meaning that pure ensemble size doesn’t have a significant impact. However, we saw that most of the improvements came from increased ensemble spread, how can we reconcile this?

Fig. 7 shows WS forecast for all C-LAEF and AROME members separately and averaged over the verification period. We can easily see that C-LAEF and AROME members are clustering together and it looks more like a multi-model EPS - this is not what we hoped for. Clustering is also observed for WG, to a lesser degree for GR, but not for RR (not shown). Although, this

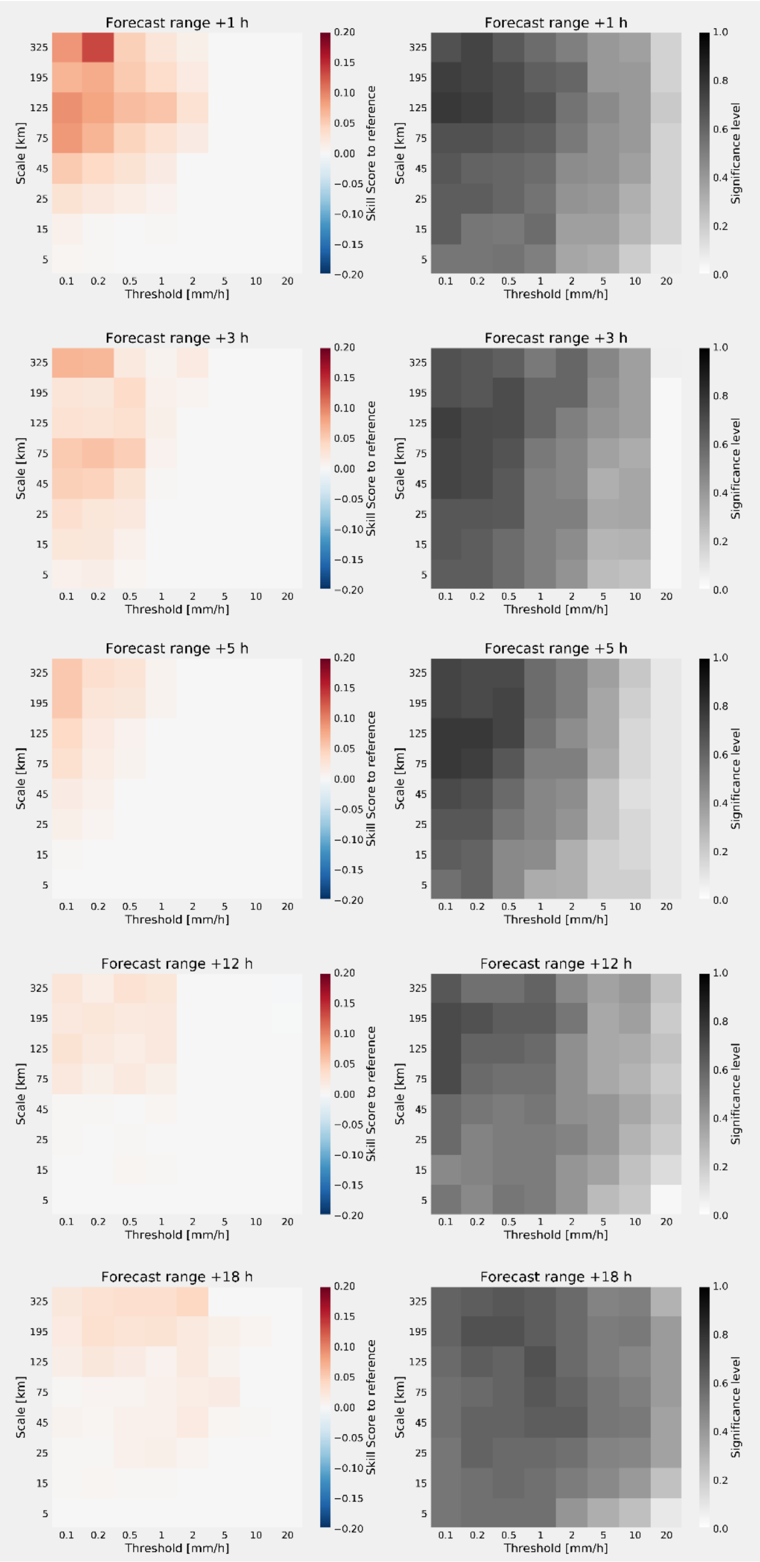

Figure 6. The left panel is the median Skill Score of FSS of LAG to FSS of REF (red means LAG is better than REF and blue the opposite) as a function of scale and threshold. The right panel is the significance level for the comparison (percentage of times FSS of LAG is higher than the FSS of REF).

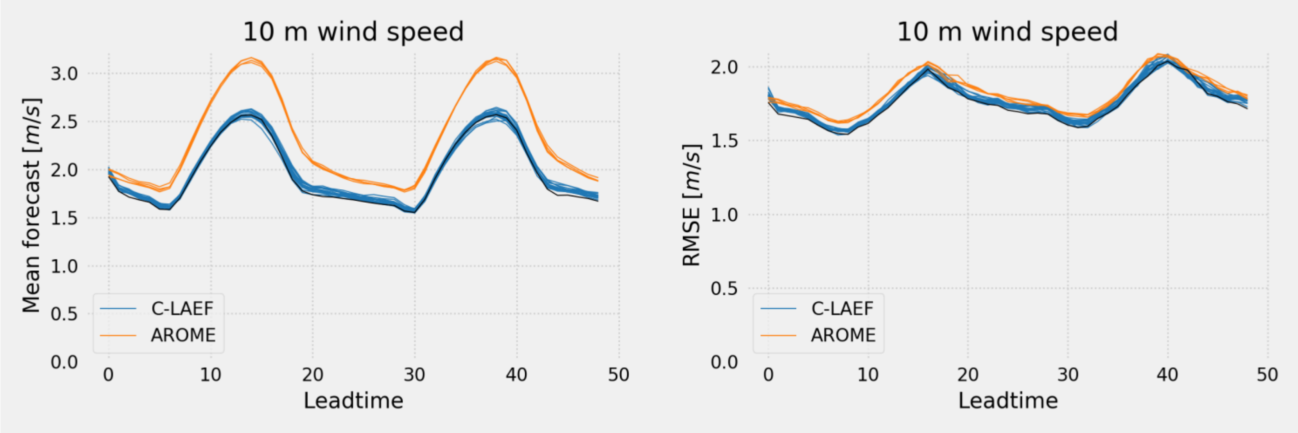

Figure 7. C-LAEF and AROME lagged members WS forecast (left) and RMSE (right) averaged over the verification period.

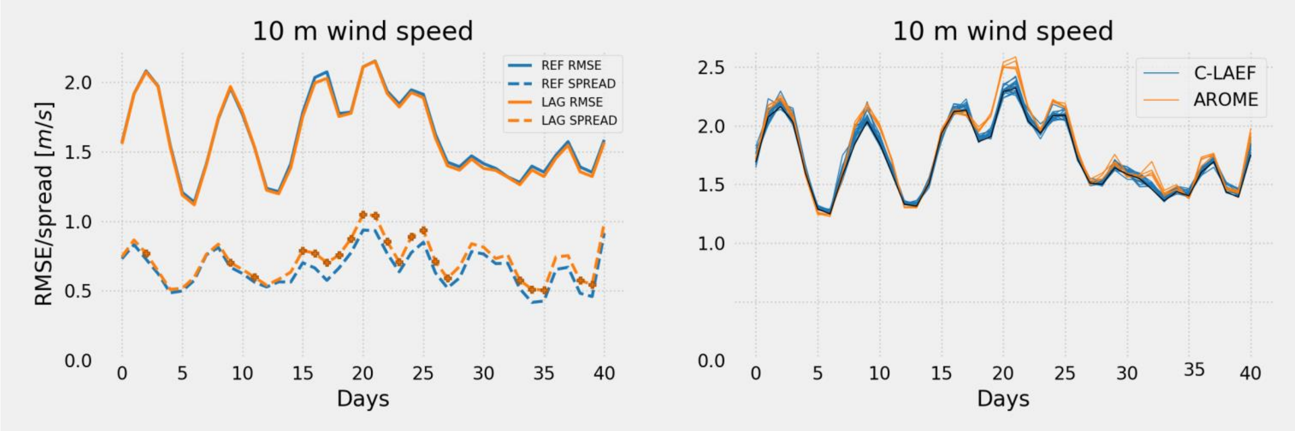

Figure 8. WS RMSE for the verification period averaged over all lead-times for ensemble mean (left) all members separately (right).

situation highly benefits the ensemble spread, members in this ensemble are not exchangeable leading to some difficulties from theoretical point of view. For example, RMSE/spread relation is not theoretically justified when members are not exchangeable; exchangeability is desirable for research ensemble forecast testing with small ensemble sizes, etc. This could be, of course, corrected by applying the same techniques used when optimally combining multi-model EPSs.

Nevertheless, from the practical point of view, things are more positive. Multi-model EPSs are known to perform very well and are in use for many years (Palmer *et al.*, 2004; Garcia-Moya *et al.*, 2011; Buizza, 2014), ECMWF-EPS just recently abandoned IC perturbation positive-negative symmetry which is known to cause clustering (Leutbecher, 2018). In addition, having a more different members is known to benefit the ensemble performance (e.g. Johnson and

Swinbank, 2009). To demonstrate those additional benefits for LAG, Fig 8 shows WS RMSE/spread for the verification period averaged over all lead-times and WS RMSE of all members separately. Although, AROME members are mostly worse than C-LAEF ones, RMSE of ensemble mean is never worse in LAG. This is the effect of error cancelation and filtering of less predictable scales when doing averages (Hagedorn, *et al.*, 2005). Even if you add a member with a higher error on average, your ensemble mean can still benefit.

What is the cause of this clustering? In the introduction we have stated that C-LAEF control is configured the same as AROME deterministic (running at ZAMG HPC). However, it is not yet 100 % the case. Some differences are intended (AROME using older GTOPO32 and C-LAEF SRTM data - this can have an impact on wind and assimilation cycles), and some were unknown up to now. For example, *LCANOPY_DRAG,* behaves differently when compiled with INTEL 16 (HPC ZAMG) and INTEL 15 (ECMWF).

Now, let’s go back to the precipitation. How can we explain the observed differences in FSS? After analysing a dozen of case studies, following conclusions emerged:

a) Improvements in LAG mostly come from two types of situations. First, C-LAEF over-predicted precipitation, while AROME members heavily under-predicted it. Although, 4 AROME members had a lower FSS compared to C-LAEF members, when all of them were added together, FSS of ensemble as a whole is improved. This is due to cancellation of biases when calculation fraction scores. Example of such cases are 31. 5. and 22. 6. Second, AROME members have a better precipitation forecast. Examples of such cases are 27. 5. and 18. 6. In former, perturbed members of C-LAEF behaved poorly, while in latter, problem was probably in data assimilation since all members of C-LAEF and AROME non-lagged deterministic behaved poorly.
b) There were some situations where adding lagged AROME members made forecast worse. For example, 10. 6. and 15. 6. where AROME members underestimate precipitation and C-LAEF ones doesn’t compensate by over-predicting it.

Conclusion

The goal of this study was to assess the impact of adding lagged deterministic model forecast (AROME) to convection-permitting LAMEPS (C-LAEF). We were hoping to get an ensemble with exchangeable members, but this was not the case due to some differences between the two models. However, practical benefits of this configuration were clearly visible in improved ensemble reliability, spread, and slightly higher accuracy for 10 m wind speed and gusts forecasts. Impact on precipitation was also positive with highest impact in the first 5 hours of integration.

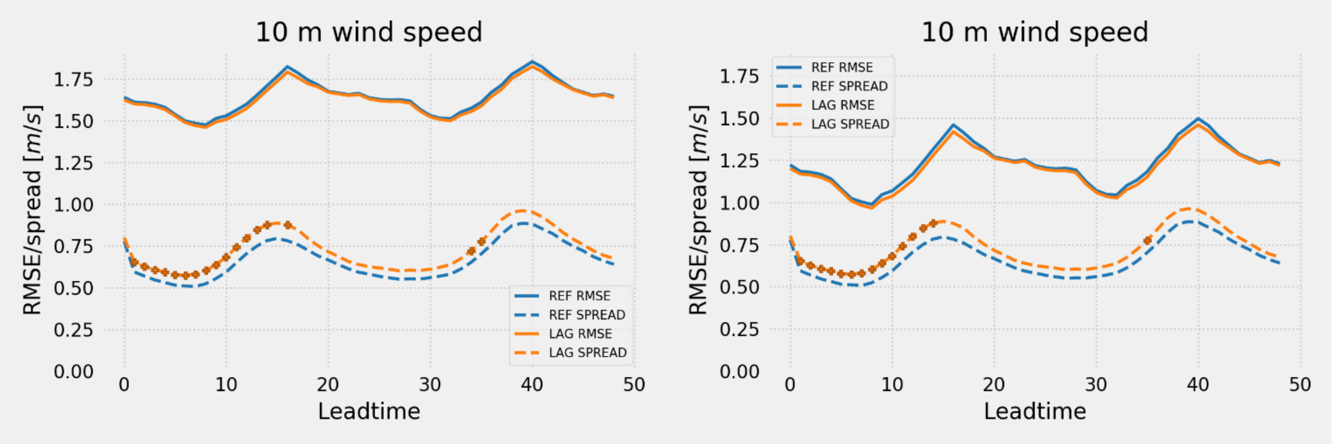

Figure 9. RMSE of the ensemble mean and ensemble spread for WS (left) and WS with observation error accounted for (right).

Last but not least, from Figs. 2 and 4, it can be seen that both experiments are heavily under-dispersive, which is a known problem in the field of ensemble forecasting and it is regularly the case in a LAMEPS studies (Keresturi *et al*., 2019). However, most of the studies, including this one, ignore observation errors. This should not be done as it makes our ensembles look worse than they actually are. Grid averaged values can not simply be compared to the point values, as representativeness error can be as high as 50 % of the total error (Haiden *et al*., 2012). As an example, Fig. 9 shows wind speed RMSE/spread from the Fig. 2 but with observation error accounted for. Now, under-dispersion is much smaller, and this is plotted for a rather small value of wind speed observation error – 1.2 m/s.

Problems in C-LAEF

As we already mentioned, two problems detected in C-LAEF are related to WG. **First**, spikes in WG speed every 6 h are present (Fig. 2.). Christoph Wittmann explored this problem further and it turned out that the differences between AROME deterministic gusts and C-LAEF gusts are due to the write-out frequency of historical files and fullpos files. In contrast to AROME deterministic (hourly ICMSH*, hourly offline fullpos) we have 6-hourly write-outs of ICMSH* and hourly inline fullpos. The reset times of gust speed is defined by the parameter *NXGSTPERIOD*=3600 - "period before write-out of historical". This is not what we want for C-LAEF. Setting *NXGUSTPERIOD*=-1 in *NAMXFU* should have the effect to reset the times for gust speed intake from the *NRAZTS* array (also used for e.g. Tmax). However, this did not have the desired effect. Finally, a logical key *LGUSTBYPOS* (introduced in cy40) was found in the code, which allows us to set the reset frequency according to the inline fullpos time-steps. Using *LGUSTBYPOS* resulted in a comparable results C-LAEF vs. AROME (in cy43 *LGUSTBYPOS* was removed again and replaced by a different key).

Second, at +000 and +001, WG speed in C-LAEF has very high error (Fig. 2). This problem is still under the investigation.

Third, a careful reader may have noticed a strange increase of WS RMSE at initial time on Fig. 9. This was a third problem detected. Partial solution came from the fact that 10 m wind was not assimilated at all. When this was corrected, control run behaved normally, but other members still had higher than expected error. Also, a deeper inspection revealed a strange behaviour of perturbed members at initial time. That is, control run (when averaged over 320 station locations) almost always has the lowest WS forecast. This definitely should not be the case as control should be contained within the ensemble. This problem is still under the investigation.

Fourth, as already discussed, *LCANOPY_DRAG* heavily depends on the Intel compiler version. This behaviour is still under investigation.

Recalculation of the V-matrix

Ensemble *Jk* method requires the, so-called, V-matrix (V) which represent error covariances of global model forecasts as described in Keresturi *et al.* (2019). Currently, C-LAEF is using an old V computed from ECMWF-EPS forecasts interpolated to C-LAEF domain by combination of GL and e927. However, configuration 903 is available and C-LAEF will be using those files for coupling and for ensemble *Jk* method in the future which means that a new V needs to be calculated.

V is modelled using ensemble simulation method where differences between 16 members of ECMWF-EPS 6 h forecasts interpolated to the C-LAEF domain are calculated two times per day (00 and 12 UTC). Due to the fact that ECMWF-EPS LBC files are only kept on MARS for about 30 days, 32-day period from 28 May until 28 of June 2019. was available for this recalculation. This gave us sample of 16 * 30 = 480 differences (two days were missing) from which to calculate V using FESTAT (cy43t2).

It is strongly recommended that V is recalculated again using additional ECMWF-EPS 6 h forecasts from different seasons to include annual variability of the model error as this can significantly affect the forecast (Storto and Randriamampianina, 2010).

While doing recalculation of the V, **a problem** of missing ECMWF-EPS BC files on MARS appeared. This was fixed by specifying “DATABASE = fdb bc” (as advised by the user support) in MARS request. However, data for 3. and 6. June were still missing.

903 bug or a feature?

In LBC files created by 903, we noticed some unexplained behaviour of wind fields with height. For our configuration, below model level 46 (about 750 hPa), orography forcing is slowly

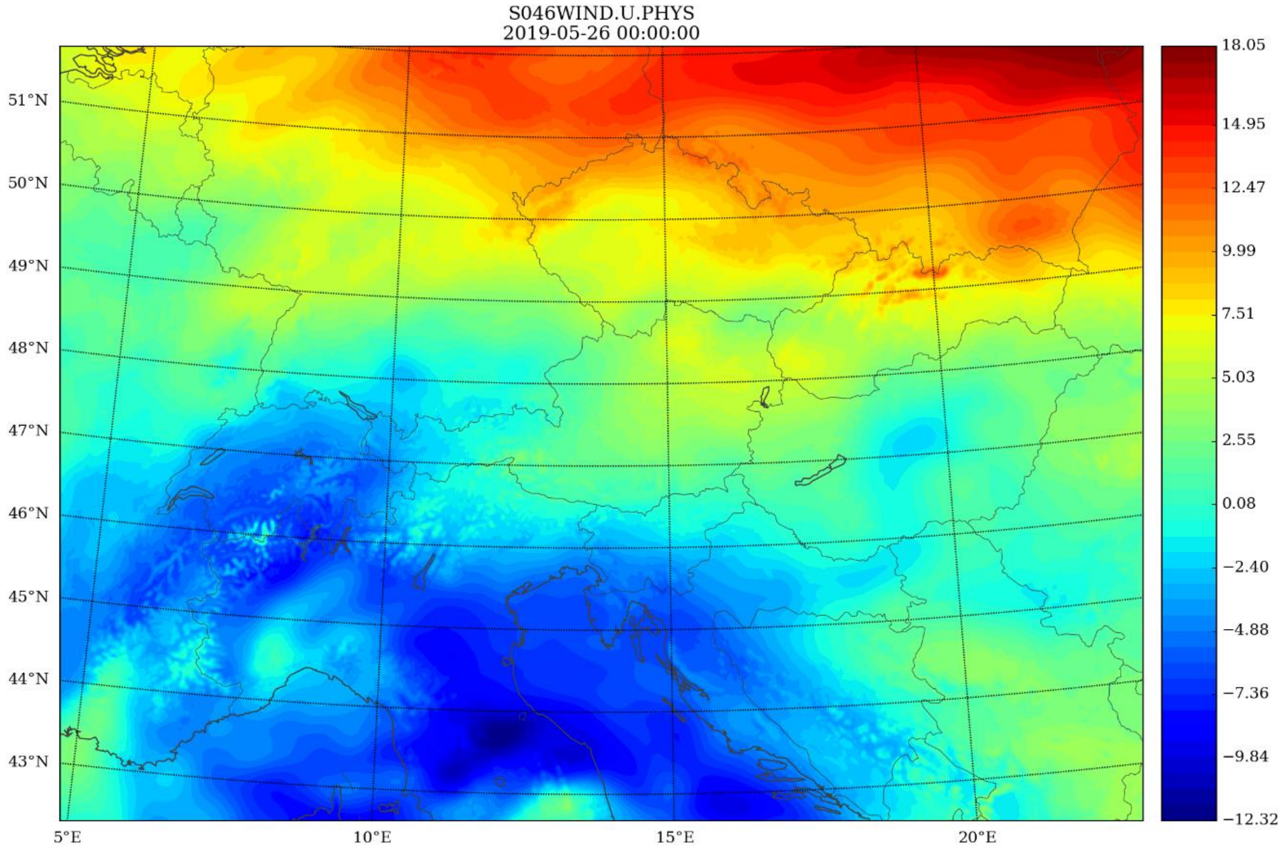

Figure 10. U-wind on level 46.

disappearing in U and V wind fields (i.e., "*WIND.U.PHYS*" and "*WIND.V.PHYS*") so that wind field looks very smooth (Figure 10-12). This is especially visible over the Alpine area. Although, fields look unrealistically smooth (e.g. Figure 12), this probably is consequence of the similar feature as in GL tool – a change between two interpolation schemes. In GL, pressure level-based interpolation is used in the free atmosphere (to preserve balances) and a terrain-following based vertical coordinate interpolation in PBL (to preserve inversions, etc.). However, this behaviour was not noticed in e927, in cy40 at least. It would be beneficial to know why is that the case.

References

Ben Bouallegue, Z., Theis, S. E. and Gebhardt, C., 2013: Enhancing COSMO-DE ensemble forecasts by inexpensive techniques. *Meteorol. Z.* **22**, 49–59.

Bowler, N. E., Arribas, A. and Kenneth, R. M., 2007: The benefits of multi-analysis and poor-man's ensembles. *Mon. Wea. Rev.*, **136**, 4113–4129.

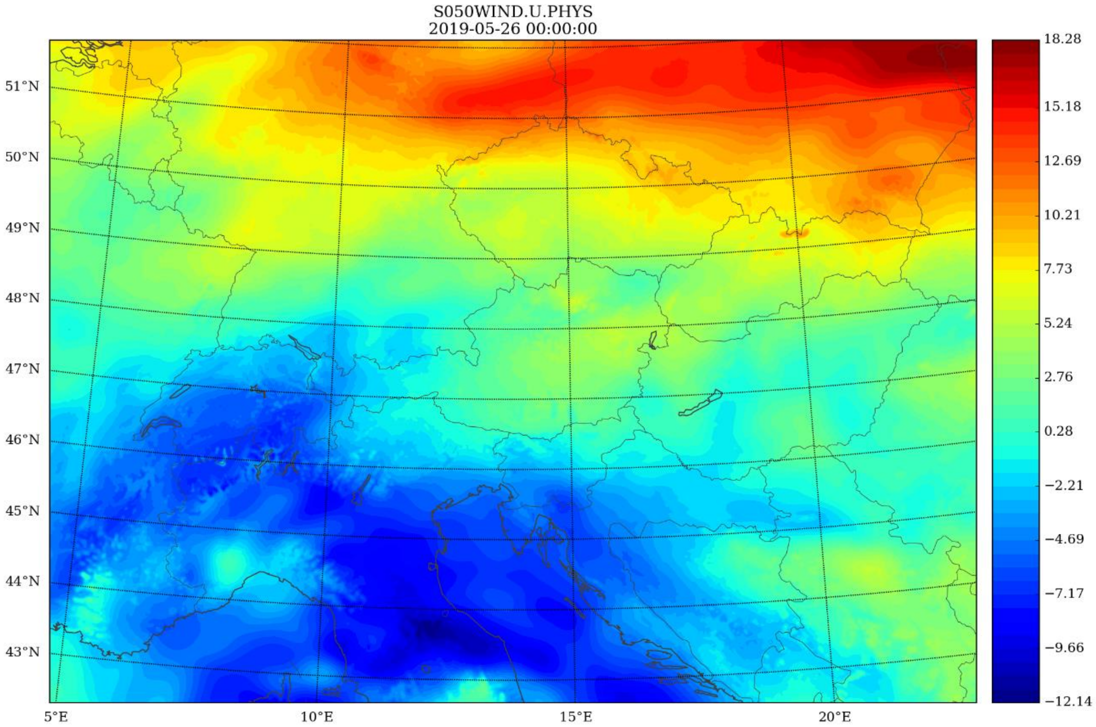

Figure 11. U-wind on level 50.

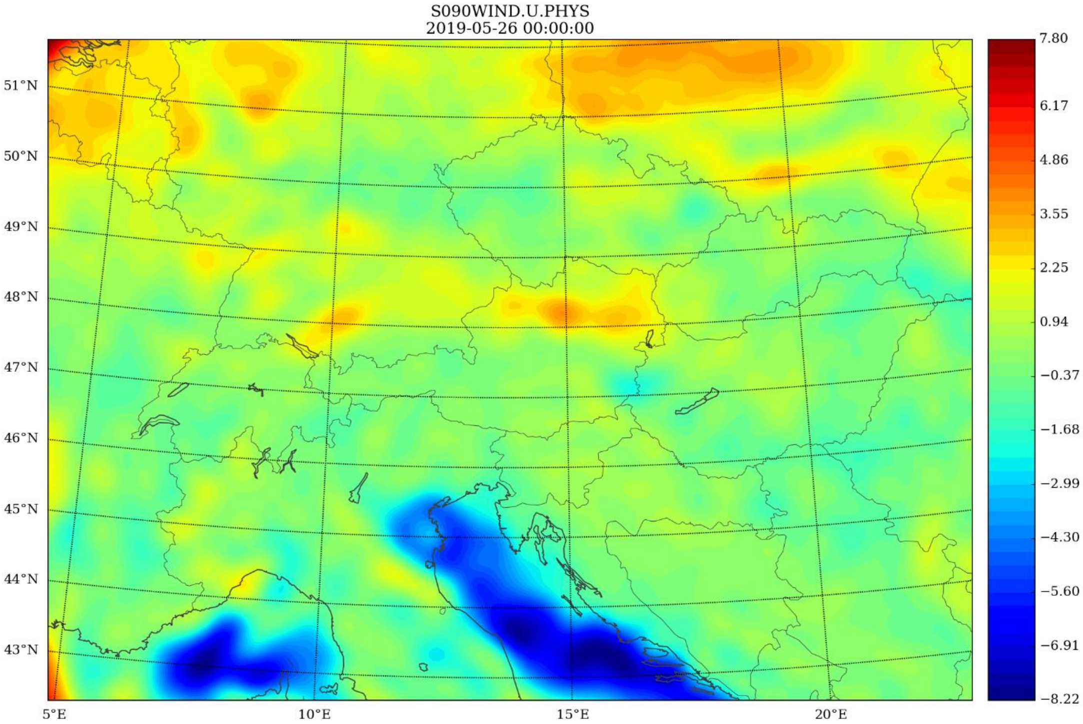

Figure 12. U-wind on level 90.

Buizza, R., 2014: The TIGGE medium range global ensembles. In: ECMWF Technical Memorandum n. 739. ECMWF, Shinfield Park, Reading, p. 53.

Duc, L., Saito, K. and Seko, H., 2013: Spatial-temporal fractions verification for high-resolution ensemble forecasts. *Tellus*, **65A**, 18171.

Ebert, E., 2008: Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework. *Meteorol. Appl*., 15, 51-64.

Garcıa-Moya, J. A., Callado, A., Escriba, P., Santos, C., Santos-Mũnoz, D. and Simarro, J., 2011: Predictability of short-range forecasting: a multimodel approach. *Tellus*, **A63,** 550–563.

Haiden, T., M. J. Rodwell, D. S. Richardson, A. Okagaki,T. Robinson, and T. Hewson, 2012: Intercomparison of globalmodel precipitation forecast skill in 2010/11 using the SEEPS score. *Mon. Wea. Rev.*, **140**, 2720–2733.

Hagedorn, R., Doblas-Rexes, F. J. and Palmer, T. N., 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept. *Tellus*, **57A**, 219-233.

Hoffman, R. N. and Kalnay, E., 1983: Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus,* **35A**, 100-118.

Keresturi, E., Wang, Y., Meier, F., Weidle, F., Wittmann, C. and Atencia, A., 2019: Improving initial condition perturbations in a convection-permitting EPS. *Q. J. R. Meteorol. Soc*., **145**, 993-1012.

Leutbecher, M., 2018: Ensemble size: How suboptimal is less than infinity? *Q. J. R. Meteorol. Soc*., 1-22.

Mittermaier, M. P., 2007: Improving short-range high-resolution model precipitation forecast skill using time-lagged ensembles. *Q. J. R. Meteorol. Soc*., **133**, 1487–1500.

Mittermaier, M., 2014: A Strategy for Verifying Near-Convection-Resolving Model Forecasts at Observing Sites. *Wea. Forecasting*, **29**, 185-204.

Palmer, T. N., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M. and co-authors. 2004: Development of a European multi-model ensemble system for seasonal to inter-annual prediction (DEMETER). *Bull. Am. Meteorol. Soc*. **85**, 853–872.

Raynaud, L., Pannekoucke, O., Arbogast, P. and Bouttier, F., 2015: Application of a Bayesian weighting for short-range lagged ensemble forecasting at the convective scale. *Q. J. R. Meteorol. Soc*., **141**, 459-468.

Raynaud, L. and Bouttier, F., 2017: The impact of horizontal resolution and ensemble size for convective-scale probabilistic forecasts. *Q. J. R. Meteorol. Soc*., **143**, 3037-3047.

Roberts, N. M. and Lean, H. W., 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev*., **136**, 78-97.

Theis, S. E., Hense, A. and Damrath, U., 2005: Probabilistic precipitation forecasts from a deterministic model: A pragmatic approach. *Meteor. Appl.*, **12**, 257–268.

Wastl, C., Wang, Y., Atencia, A. and Wittmann, C., 2019: Independent perturbations for physics parametrization tendencies in a convection-permitting ensemble (pSPPT). *Geosci. Model Dev.*, **12**, 261–273.