Minutes of the meeting of COSMO WG7 (WG on Predictability and Ensemble Methods) and PROPHECY PP, Web-conference, 18th and 19th March 2021.

Participants:

Dmitry Alferov (RHM), Marco Arpagaus (MCH), Evripidis Augoustoglou (HNMS), Elena Astakhova (RHM), Sascha Bellaire (MCH), Anastasia Bundel (RHM), Grzegorz Duniec (IMGW), Valeria Egorova (RHM), Felix Fundel (DWD), Dmitry Gayfulin (RHM), Christoph Gebhardt (DWD), Flora Gofa (HNMS), Pavel Khain (IMS), Daniel Leuenberger (MCH), Yoav Levi (IMS), Francesca Marcucci (COMET), Chiara Marsigli (DWD, WG7 coordinator), Andrzej Mazur (IMGW), Dmitrii Mironov (DWD), Christoph Schraff (DWD), Martin Sprengel (DWD), Michael Tsyrulnikov (RHM), Andre´ Walser (MCH).

Not all the participants took part to the whole meeting.

Minutes:

Thursday 18th of March 2021, from 13:30 to 16:30

The meeting alternates presentations and discussion.

- Chiara Marsigli: Parameter list for ICON and ICON-LEPS.
The list of ICON parameters suitable for tuning and model perturbations is available on the cosmo website (e.g. link from the WG7 page). The COSMO members are invited to report here their experience about the usage of these parameters.
The transition from COSMO-LEPS to ICON-LEPS has been approved and will be performed next year. A task of PROPHECY is the test of the model perturbation in ICON-LEPS. As preparatory step to this work, this year E. Augoustoglou will perform a sensitivity test of ICON to parameter variations over a Mediterranean domain.

- Evripidis Augoustoglou: "Status of ICON implementation at ECMWF to study parameter sensitivity"
The help offered by IMS colleagues in the implementation and run of ICON is warmly acknowledged.
The simulations will be made using ICON R3B7 (about 6.5 km resolution). It is discussed if to used a higher resolution (5 km) set-up but the need of sticking to the future ICON-LEPS configuration determines the choice of the 6.km resolution. On top, CS comments that ICON has not been tested enough at 5 km.
It is proposed to extend the domain to include Israel or to run two separate domains to cover as much as possible the eastern Mediterranean.
Decision: The decision is [email exchange after the meeting] to run on one domain, approximately the one defined by I. Cerenzia, at 7 km resolution. It will be explored the possibility, as suggested by Y. Levi, to extend the domain include the common verification area.

- Flora Gofa, Anastasia Bundel and Chiara Marsigli describe the proposal for a follow up of the AWARE PP, which will focus also on object-based verification for ensembles.
The possible interest of WG7 members is asked. No comments are made.

Decision: the definition of the proposal will continue by email, in order to coordinate the activity with WG7, where possible.

- Andrzej Mazur: "PROPHECY activities at IMGW - bow echo / derecho case study of August 11th, 2017"
A series of nested runs with COSMO at 7, 2.8 and 0.7 km are performed over Poland to investigate the performances on a severe weather event. It is tested the effect of switching off the parametrisation of the convection.

- Pavel Khain: "The effect of SPPT and parameters perturbations on the ensemble quality"
It is shown the positive effect of the combination of SPPT and PP, particularly for winter and autumn.
It is also shown the impact of the cluster analysis of the Boundary Conditions: the random choice seems to lead to the best results.
Discussion:
C. Gebhardt suggests to consider the resolution component of the Brier Score and not only the ROC area, since the ROC curves look very close together and their area seem to be affected by how the interpolation between the points of the curve is performed. It is suggested that the ranking of different the clustering methods depend too strongly on details of the verification implementation.
M. Tsyrulnikov suggests to compute the statistical significance of the scores and to consider if the results are dependent on the weather situations.
E. Astakhova asks if they considered also the impact of PP alone, without SPPT. This has not been tested.
P. Khain asks is the different perturbations can interfere leading to a partial cancelation of their effect. C. Marsigli answers that this effect has been reported by several people, e.g. at EWGLAM.
M. Arpagaus asks the settings of SPPT and PP: MCH settings for SPPT and DWD for PP, with slight modifications. For clustering, it has been tested also the original set-up used in COSMO-LEPS, which is similar to the one implemented at MCH (where T instead of Z is used). No weighting of the representative members of the clusters is performed, neither in COSMO-LEPS nor at MCH, while in the IMS tests this has been adopted.
C. Schraff asks how the spread-skill relation is computed: according to the paper by Klasa et al (2018).

In PROPHECY three groups are working to include model perturbations in ICON: SMME at DWD, AMPT at RHM and (i)SPPT at MCH. Their contributions are being coordinated in order to propose a unique approach for the three implementation in the model, if possible.

- Elena Astakhova, Mikhail Tsyrulnikov: "Towards the SPG/AMPT model-error generator in ICON-LAM"
An update is provided on the strategy to implement the AMPT scheme in ICON. Initially the idea was to perturbed the prognostic fields, then for reasons of compatibility with the code structure it was decided to follow the same approach of M. Sprengel for SMME, positioning AMPT as an additional slow-physics parametrization scheme and perturbing the diagnostic fields (T, u, v). It is still open the issue of how to perturb the hydrometeors.
Next steps will include also: speed up the SPG, by trying a new FFT package (FFTW), and revisit the SPG vertical length-scale specification.
Discussion:

M. Arpagaus underlines how for MCH it is crucial to find a method for perturbing the moisture variables, which play a major role for their ensemble.
M. Sprengel did not try as well to perturb the moisture variables (his focus was in the variables for renewable energy), but he thinks that this was tried by T. Heppelmann, who therefore likely found the appropriate place in the code where this can be done.
S. Bellaire says that for MCH the strategy proposed by M. Sprengel and M. Tsyrulnikov is also fine, at least for the part of the slow physics tendencies. However, they have to find another solution for the other tendencies.

- Christoph Gebhardt: "ICON-D2-EPS: status and plans"
ICON-D2-EPS is operational since the 10th of February 2021. The scores are good but there is little spread in the wind, also in the gusts. ICON has just been upgraded (not in the operational version, yet) with the cp/cv bugfix + grayzone tuning + ecRad radiation scheme + perturbations of gkwake and gfrcrit in the ensemble. There was a bug in the implementation of the ensemble due to a problem in the RNG, now it is going to be fixed as well.
Plans: test stochastic parametrisations in ICON-D2-EPS, Stochastic Shallow Convection (implemented in ICON by M. Ahlgrimm) and SPS2 (developed at LMU).

<u>Friday 19th of March 2021, from 9:30 to 12:30</u>

The meeting starts with a discussion on model perturbation.

Decision: complete the implementation document based on the proposal by M. Sprengel, revised by M. Tsyrulnikov. Next week the document will be completed and then sent to D. Rieger and G. Zängl. It will be underlined in the document that the strategy should provide a solution also for the perturbation of the moisture variables.

On the WG7 page there is now a section on the Stochastic Workshop, hosting almost all the presentations. The minutes of the discussion will come soon.

After a question by P. Khain, it is recalled that the members of the IFS ENS are constructed with the aim to be probabilistically indistinguishable, with the exception of the control member. In this sense, to select the first 20 members is similar to a random choice. CG adds that this is not the case for the ICON-EPS members, where the perturbations are fixed and depend on the member ID.

The joint session WG7-WG1 takes place, dedicated to the spread-error estimate.

C. Marsigli introduces the topic with a short presentation, recalling that the ensemble spread should match the RMSE of the ensemble mean, statistically over a long enough period. This is derived under the hypotheses of the members being interchangeable, no bias in the model, and a large enough ensemble size.

M. Arpagaus invites to use for this purpose the error standard deviation instead of the RMSE, to get rid of the model bias. Anyway (CS,CM) this does not permit to get rid of the conditional biases.

C. Schraff presents a work on the introduction of the observation error in the spread-error ratio.
A comparison is made between the metrics

SSR1=spread/SD(O-FG) and
SSR2=sqrt(spread**2+obserr**2)/SD(O-FG)
where obserr includes instrumental error, representativeness error and observation operator error.
When SSR1 is used, the ILAM ensemble has lower ratio than the COSMO ensemble (more underdispersive), when SSR2 is used ILAM ensemble has higher ration of the COSMO ensemble and for some variables is even overdispersive.

Discussion:
A. Bundel asks is also the random error of the observations is included in obserr
F. Fundel asks if it is possible to add the obserr contribution also in terms of standard deviation. Answers by CS is that the equation is also valid in terms of variances.
F. Fundel: the observation operator error is part of the model error, it does not belong to the observations, therefore should be accounted for in the bias and not in the observation error
C. Schraff: it can be considered where to add it, but anyway should be considered
[comment: and should not be counted twice]
C. Schraff: the estimate of the observation error is taken from the Desroziers statistics, providing already the contribution of the 3 terms. It provides a correct estimate of the observation error if a correct estimate is provided in input to the filter. If the spread is wrong, the observation error estimated by Desroziers is also wrong, but this problem can be solved with an iterative process: the system converges fast to a correct estimate, independently from the information at the start. In KENDA the spread-error relation is measured by the *rho* parameter. If the spread is too small, *rho* is greater than 1 and gives inflation, and viceversa. It is limited between 0.5 and 3. In KENDA it is often observed a too large spread, leading to deflation.
C. Schraff: in the computation of SSR2 different values of obserr have been tried (without changing the value in the cycle itself, only diagnostically).
- When obserr is decreased, rho increases, the spread-error ratio increases, O-A decreases because more weight is given to observations (less weight being given to the first guess). But, as a consequence, the errors of the next first guess (after 1 hour of forecast) increase slightly.
- When obserr is increased, rho decreases, the spread-error ratio decreases, O-A increases because less weight is given to observations (more weight being given to the first guess). But, as a consequence, the errors of the next first guess (after 1 hour of forecast) decrease slightly.
It is noted by M. Arpagaus and C. Marsigli that this effect seems to be a bit surprising.
M. Tsyrulnikov describes the effect in this way: from the model point of view the observations have too much variability because they contain measurement error and representativness error (both absent in the model forecast), therefore the forecast-minus-obs RMS should be reduced (decrease RMSE), while from the observation point of view the model has too little variability because it lacks subgrid scales and measurement error (increase ensemble spread).
M. Tsyrulnikov: the observation operator error should be considered a representativness error.
M. Tsyrulnikov comments about the Desroziers statistics, stating that it relies on a number of assumptions. In particular, it postulates no cross-correlation between obs-error and forecast error. This assumption may be violated with satellites and radars that have spatially and temporally correlated errors. So, conclusions from the Desroziers statistics should be taken with care.

F. Fundel gives a presentation about representativness.
The different nature of model and observations introduces the need, for their comparison, of quantify the representativness of the observations on the model grid scale.

An approach has been proposed by Ben Bouallegue (2020, ECMWF Technical Memorandum Nr. 865): in order to include the effect in the spread-error evaluation, an approach based on parametric models has been developed (normal distribution is used for 2mT while censored shifted gamma distribution is used for precipitation). Model fitting is performed with pairs of observations and area averages (for various scales), optimization being made with CRPS. Verification of the ensemble is then performed with a "perturbed ensemble approach", where observation uncertainty is added on top of the ensemble uncertainty (processing of the ensemble before the score computation).
A problem in this approach for our applications is in the grid scale size: usually we do not have enough observations is a box (at least 5). Results were shown for ensemble verification only, while the method should also be applicable to the deterministic verification.
Another method has been proposed in the ICCARUS talk of P. Khain, where a geostatistic approach is adopted. When gridded observations (from INCA) are obtained (e.g. with a Kriging), an estimate of uncertainty is derived, which can be used to "correct" the scores. A disadvantage relies in the fact that this approach is possible for some parameters in the domain of the data-set only and only if the data-set is on the model grid scale.
F. Fundel proposes to explore the usage of the difference between verification against observation and against analysis to estimate representativness error (OBS-ANA verification). In the feedback files the analysis verification is also provided, on the observation points (in the observation space). This gives the advantage to have an estimate of the representativness error available for all kind of observations that is already included in the feedback files (easy to integrate in a verification).
Summarising: at long lead-times (at the beginning is too much the influence of all possible kind of errors), can the difference of observation versus analysis verification provide an estimate of observation representativness?
Discussion:
D. Leuenberger and M. Arpagaus the spatial variability of the representativness error is high, particularly in complex terrain, and depends on the geographical area.
F. Fundel: here a correction factor is proposed, to be applied to the score computed over the whole domain, therefore it is not meant as a point estimate.

The proposed methods are considered very interesting. It is proposed to continue the exchange and the discussion on this issue in the next meetings.