Review of gmd-2016-191

Title: On the forecast skills of a convection permitting ensemble

Authors: T. Schellander-Gorgas, Y. Wang, F. Meier, F. Weidle, C. Wittmann, and A. Kann

General comments:

This manuscript examines 16-member 11- and 2.5-km ensemble forecasts over a 3-month summer period focusing on convection over Austria. Most of the evaluation regards verification of probabilistic precipitation forecasts at fairly light precipitation thresholds. A variety of verification metrics are appropriately used.

Overall, the manuscript is well written. Although similar material has been explored elsewhere, I think the topic and novelty is nonetheless sufficient to warrant publication of this work. In my opinion, only some minor revisions are needed.

Bigger comments:

1. You did not cite or discuss Duc et al. (2013), which is highly relevant to your work, as they examined 2- and 10-km ensembles. Their conclusions were broadly similar to yours. I suggest briefly discussing Duc et al. (2013) in page 4 lines 5-9, and throughout, pointing to similarities between your work and theirs. Schwartz et al. (2009) might also be worth mentioning at times, but citing Duc et al. (2013) is more critical.

2. I question the need to show the ensemble mean curves on Fig. 5. The ensemble mean, as you later note, is smooth and unrealistic for heavier rainfall rates. Can the curves for the ensemble mean simply be removed? Overall, you could be more precise in the text about when you're showing curves for the mean (as in Fig. 4) versus the members.

3. I believe section 4.2.1 about the Brier score (BS) is incomplete and potentially a little misleading. I think that rather than showing the BS, which depends on the observations (the uncertainty term), that showing reliability and resolution explicitly is more beneficial, as some of the behaviors you noted are quite likely due to the uncertainty term dominating. Also, I noticed you listed in Table 2 "reliability", "resolution", and "uncertainty" but never discussed them in the text.

Smaller comments:

1. Page 3, lines 1-5: What's the difference between "convection permitting" and "convection allowing"? Do you mean them synonymously?

2. Page 3, line 21: Schwartz et al. (2015) is a better reference for a real-time NCAR convection permitting ensemble system than Schwartz et al. (2014). Suggest making this change.
3. Page 4, lines 7-9: Not sure how this sentence follows from the previous one or is relevant. Suggest omitting and instead discussing Duc et al. (2013).
4. Since ALADIN-LAEF used mixed physics, is it fair to treat the members as being equally likely? Any comments on this?
5. Page 7: It should be section "2.3" not "3.2".
6. Page 8, line 30: So you were using a block-bootstrapping approach? How did you settle on a block length of 8? Also, to what does 8 refer? 8 forecast hours?
7. Page 9, Eq. (2): Please be more precise about x_i, which is 1 if the event occurred, and 0 otherwise.
8. Page 10, Eq. (7): Why are there overbars on R?
9. Page 13, line 20: Suggest "…the forecast probabilities **and** observed values."
10. Page 13, line 22: What do you mean by "signals of CRPS"?
11. Page 13, lines 30-31: Suggest "…an improvement for bias and CRPS at a significance…"
12. Page 14, lines 13-14: Fig. 5e,f don't fully support this statement.
13. Page 14, line 19: I don't believe this statement is fully correct—AROME in Fig. 5b reaches its maximum at 1800 UTC.
14. Page 15, lines 8-10: Please rewrite the beginning of this sentence to make it clearer.
15. Page 15, line 10: Can you perhaps add a brief concluding paragraph summarizing the main points of Fig. 5?
16. Page 16, line 14: What do you mean by "on a low level"?
17. Page 17, line 16: "FSSs" not "fractional skill scores".
18. Page 17, section 4.2.3: Might want to note that your results are quite consistent with Schwartz et al. (2009) and Duc et al. (2013).
19. Page 17, line 23: Don't think "reliable" is the right word.
20. Page 17, line 26: Is "exemplarily" the right word?
21. Fig. 4: The line labels for AROME and ALADIN should be enlarged. Also, please note in the caption and text that these statistics are for ensemble means. Finally, please note the units either in the y-axis labels or figure caption.
22. Fig. 9: What do the shadings mean? Suggest the first line of the caption reads as "…between the centers of mass of **observed** precipitation objects…"
23. Fig 10: What do you mean by "averages" in the caption? Were the statistics aggregated or averaged? Why sum over all times rather than showing a time-series? Also, please change the beginning of the caption to "FSS" rather than "fractional skill scores".
24. Fig. 11: The colorbar should be bigger and possibly just in one location.
25. Fig. 12: Does the 3rd line of the caption describing the shadings apply to

both (a) and (b)? Also, why are you showing the ensemble means in (c) when in Figs. 7 and 8 you showed data from individual members?

References:

Duc, L., K. Saito, and H. Seko, 2013: Spatial–temporal fractions verification for high-resolution ensemble forecasts. *Tellus*, **65A**, 18171, doi:10.3402/tellusa.v65i0.18171.

Schwartz, C. S., and Coauthors, 2009: Next-day convection-allowing WRF Model guidance: A second look at 2-km versus 4-km grid spacing. *Mon. Wea. Rev.*, **137**, 3351–3372, doi:10.1175/2009MWR2924.1.

Schwartz, C. S., G. S. Romine, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2015: NCAR's experimental real-time convection-allowing ensemble prediction system. *Wea. Forecasting*, **30**, 1645–1654, doi:10.1175/WAF-D-15-0103.1.