We thank Pekka Rossi for the useful remarks and also for spotting some mistakes in the paper (reference+equation). Below you will find the answers to your questions.

1. Introduction: I think it is quite widely accepted that the term “nowcasting” refers to very short range forecasting in the time range 0-6 hours (e.g. https://www.wmo.int/pages/prog/amp/pwsp/Nowcasting.htm). To avoid confusion, it would be better to first define the term nowcasting with this definition, and then specify that this paper considers only the first two nowcasting hours.
Good idea. We will define the term nowcasting in the introduction and mention that we only focus on the first two hours. Nowcasting is strongly based on the use and extrapolation of real-time observations. During recent years there has been significant progress in NWP modelling with radar data assimilation techniques, which reduces the length of the nowcasting time range (e.g. the AMS glossary of meteorology cites 3 hours). In the future we expect to have a seamless transition between the observations, empirical and NWP forecasts, which will make the definition of the nowcasting time range even fuzzier.

2. Introduction: The paper underlines advantages of radar-based nowcasting over NWP during the first nowcasting hours. To be fair, authors should better acknowledge that NWP typically outperform radar-based nowcasting after a few forecast hours, which is still in the nowcasting time range (assuming that the definition of 0-6 hours is adopted). I also think the paper should acknowledge that NWP community working very hard to improve the nowcasting of rainfall (see e.g. Sun, J., and Coauthors, 2014: Use of NWP for Nowcasting Convective Precipitation: Recent Progress and Challenges. Bulletin of the American Meteorological Society, 95, 409-426.).
Exactly, NWP can already provide useful skill over extrapolation techniques after 2-3 hours lead time. We will cite the progress in the NWP community, but also the practical fact that rapid-update (5-10 minutes) high-resolution radar-data assimilating ensembles in the nowcasting range are not yet a reality nowadays. Thunderstorms can exhibit significant evolution over a few tens of minutes and the current operational NWP systems are not able to reach the resolution, update frequency and skill of empirical nowcasting techniques in the first couple of hours.

3. The verification was performed with four case studies. This is not very extensive verification, given the availability of radar data and low computational costs nowadays. I do not feel strongly enough about this to make it a major issue, but perhaps Authors could underline that more extensive evaluation would be needed to capture full performance of the system.
Yes, indeed we verified only 2 convective and 2 stratiform precipitation cases. The deterministic verification can give quite different results depending on the cases. More data would be needed to better highlight the climatological spatial distribution of STEPS forecast errors (e.g. as done in Foresti and Seed, 2015). On the other hand, the probabilistic verification converges much faster to stable statistics. This is due to the fact that the probabilistic verification pools the data in both space and time, which gives many more samples to compute the statistics.

4. I was a bit surprised that Authors did not do any comparison against a reference system (e.g. basic deterministic extrapolation). It would have been interesting to see differences between a legacy system and STEPS-BE (e.g. in terms of RMSE, GSS).
Indeed, it could have been an additional analysis. In Foresti and Seed (2015) we presented a comparison of the STEPS ensemble mean and the Eulerian persistence forecast. The STEPS ensemble mean was better most of the time except for the regions with reduced visibility due to orography and far from the radar. Comparing the ensemble mean forecast to a deterministic control forecast in terms of pixel-based RMSE would reward the ensemble mean. In fact, the smoothing effect due to ensemble averaging filters out the unpredictable features and reduces the double penalty error occurring when

forecasting for example a storm at the wrong location. Comparing the probabilistic forecast error of STEPS against the probabilistic error of a single deterministic forecast (issuing only 0 or 100% probability of rain) also suffers from the dependence of scores with the ensemble size. In fact, larger is the ensemble size smaller is the Brier score (Ferro, 2007), which complicates the comparison of ensemble prediction systems composed of different members (20 for STEPS and "1" for the deterministic control forecast). We could have done a comparison to show that STEPS is better than a deterministic nowcast, but a fair comparison that considers the influence of ensemble size and the different statistical properties of the competing forecast models (e.g. smoothness) would have been much more complex.

5. P. 6850: "Another explanation for this underestimation is due to not using a model for the radar measurement errors, in particular due to the space– time variability of the Z–R relationship". It is not clear to me how errors due to initial conditions can be observed in this verification, because the reference data applied in the verification data is obtained from the same erroneous radar data.
We will add some sentences to better formulate this concept. The last observed rainfall field is extrapolated using a fixed Z-R relationship. The same Z-R relationship is used to convert the observed reflectivity to the rainfall rates that are used for the verification. However, spatial and temporal changes in the drop size distribution (DSD) can lead to changes in the estimated rainfall rate that is used for the verification. Therefore, there could be a mismatch between the "fixed" DSD of the forecasts and the variable DSD underlying the verifying observations.
Another possible source of mismatch could be due to the advection correction with optical flow. The forecast accumulations are computed by advecting forward the previous rainfall field. On the other hand, the observed accumulations are computed by reversing the optical flow vectors and advecting the rainfall field backwards. This choice increases the differences when comparing the +0-5 min forecast accumulations (advection of the "0" min image forward) with the +0-5 min observed accumulations a posteriori (advection of the "+5" min image backwards). We will add these details to the text.

6. Authors might want to revise the use of the term skill. Isn't it by definition a measure of forecast accuracy with respect to the accuracy of a reference forecast? The term is quite widely used throughout the text.
Good remark. We will revise the text and replace the term skill with a more appropriate one to be consistent with the terminology used in forecast verification.

7. P. 6587 and p. 6849 (Brier score and Brier skill score), also related to my previous comment. Brier score (BS) is a measure of accuracy, and BSS compares BS of two systems. Thus, I believe it would be better to say that "The Brier skill score characterizes the relative accuracy of the probabilistic forecast compared to a reference system". Although climatology or sample climatology is often used as a reference, BSS can also be computed against other reference forecasts, e.g. another probability forecasting method or even a deterministic forecasting method treated as a probabilistic binary forecast.
Thanks. We will specify that the reference can be different than the sample climatological frequency.

8. P. 6858: Foresti et al. (2013). I couldn't find it in the reference list. Foresti et al. (2014)?
Well spotted. It is the paper on the analogues written in 2013 but published in 2015.

9. eq. (8). It seems that index m is not defined. Shouldn't the index i under the square root be replaced with m?
Thanks for remarking the typo in the equation. The second summation is done from m=1 to M.

References

Ferro, CAT. (2007) Comparing probabilistic forecasting systems with the Brier score. *Wea. Forecasting*, 22:1076–1088.

Foresti, L., and Seed, A. (2015) On the spatial distribution of rainfall nowcasting errors due to orographic forcing. *Meteorological Applications*, 22(1):60–74.