Anonymous Referee #1

General comments:

The paper entitled "A flood episode in Northern Italy: multi-model and single-model mesoscale meteorological ensembles for hydrological predictions" by S. Davolio et al. presents an investigation of uncertainties in the meteo-hydrological forecast chain and sources of these uncertainties. Specifically, the focus is on the differences of precipitation forecast of LAM ensembles with different ensemble generation strategies and the effect on discharge forecasts by a rainfall-runoff model (also in comparison to the results based on a global model NWP EPS as input to the discharge model). This investigation is carried out on the basis of a flood episode in Northern Italy with two events affecting the Reno River (Apennines).

The topic of the paper is up-to-date and addresses relevant questions. Most of the uncertainty in such approaches of river discharge predictions has its origin in the meteorological input. Therefore, investigating different aspects of this NWP uncertainty and their effect on the discharge forecast is essential for dealing with probabilistic approaches in hydrological modelling. This approaches and results are important contributors to decision making processes.

The paper contains a good description of the state of the art. The applied methods are well described, the structure is clear and the road-map of this investigation is easy to follow both from a scientific point of view as well as concerning the quality of the presentation. The study is based on sufficient forecast data of state-of-the-art model and discharge observations. The caveat related to general conclusions based on case studies is included in the paper. Even though the paper is based on one case study, the results are sufficient to highlight important aspects of the problem and to trigger further research.

We would like to thank Referee #1 for the helpful suggestions. In particular, the latter comment provides a good description of the typology of our study and is suitable and helpful for answering the major criticism raised by Referee #2. As Referee #1 properly states, we have honestly stressed the main limiting aspect of the present study, i.e. the fact it is based on one case study only (page 13434). The purpose of the paper is to show that the information conveyed by a multi-model ensemble allows to properly address the potential threat associated with the single case study discussed here and to underline the main issues which should be addressed by future research. Within this context, we tried to analyse in detail the behaviour of the multi-model ensemble, discussing the peculiarities of these systems for this case. Thus, this paper is supposed to present "just" the starting point of a possible long-lasting research we have planned, whose motivation is better described in the following answer, and to foster further research.

The language is good (as far as I can judge this as a non-native speaker). However, needs some minor revisions concerning motivation for and scope of the paper. Furthermore, one aspect of the results deserves more attention. Those aspects are briefly described below together with a list of additional minor comments.

I can recommend the paper for publication after some work has been done on those minor aspects.

A) Scope of the paper / motivation: The state of the art and related literature is actually very well described in the introduction. However, the statements relating this to the specific research target and the scope of this paper have to be more precise and significant. For example, the sentence containing the statement "..two different. . .approaches. . .are compared" is too weak, but currently, there are no stronger statements of motivation for the work which has been done. Statements like: "Considering the entire meteo-hydrological chain, the lack of theoretical development supporting strategies for flood forecasting leaves room for testing ad hoc methodologies on a case by case basis (Cloke and Pappenberger, 2009)." sound a bit like "We are doing it, because according to a review paper, we can try anything". It would be good to have a few more statements on why the

authors did exactly what they did (why single-model versus multi-model). This can be easily done by distinguishing their work from i.e. those by Adams & Ostrowsky (forecast range) and Addor et al. (single model EPS) so that the reader knows what is really new and can easily get an idea of the overall research target of the presented work. Furthermore, the authors could motivate their work with better arguments, e.g. referring to the "meteorology-only" literature or the other aspects of review papers like Cloke & Pappenberger and Cuo et al. All this is already hidden in the current text but needs some more depth in the argumentation to highlight those aspects of the work which bring an additional contribution to meteo-hydrological forecasting.

We agree with the Referee. In particular, the sentence "*the lack of theoretical development…"* can be easily interpreted in the suggested sense. We will explain more clearly the motivation for the work in the reviewed paper as follows. An ensemble system (COSMO LEPS) has been already operational at ARPA-SIMC for several years and coupled with TOPKAPI in order to provide operational discharge predictions for civil protection purposes. At the same time, collaborative research activities have been carried on between ARPA-SIMC and CNR-ISAC concerning NWP model applications to hydrological forecasting (Diomede et al., 2008; Davolio et al., 2008), aimed at testing different models, resolutions, analyses. From the one hand, previous studies (Marsigli et al., 2008) suggested the possibility of improving the statistical performance of the operational forecasting system based on COSMO LEPS. On the other hand, the availability of different state-of-the-art limited area models, employed in the two institutes for real-time forecasting and constantly updated and validated, allows to implement a multi-model meteorological system. Within this context, our aim is to answer the following question: is it possible to improve the performance of a single-model ensemble (the same implemented by Addor et al., 2011), also in terms of hydrological predictions, using the information that can be "easily" obtained by a multi-model system? The starting point in order to answer this question is a comparison between the two ensemble modelling systems, to identify possible pro and cons, for a single event, looking not only at the short range (as in Adams & Ostrowsky, 2010), but also at longer lead time. A case study approach does not complete our investigation, but, as we tried to point out in the conclusion, represents just the starting point of a long and complex task.

B) Extra from hydrological model

Beside mentioning the spread (e.g. represented by the 10th and 90th percent quantiles), section 5 could put more emphasis on the interpretation of the best discharge model forecasts (in this case study) as scenarios which provide decisive information for the comparison of the EPS approaches and their effect on river discharge forecasts as well as for any potential decision making based on such forecasts. The focus right now is on the spread which is of course influenced by the "extreme members" but at the time of the observed peaks, those "extreme members" can make the difference between the multi-model EPS and the COSMO-LEPS. E.g. for the first peak and the investigation of the shorter forecast range (bottom row in Fig. 6) the distance between the 10th and 90th percent quantiles is not significantly different for the two EPS approaches, however the most extreme member of the multi-model EPS provides the decisive information for the warning level which is not revealed even by the 10/90th percent quantiles. Such effects cannot be seen in the meteo-only forecasts if looking at exceedance probabilities as it is done in Fig. 4 and 5. This is an important aspect of the investigation of uncertainty in such a meteo-hydrological forecast chain.

It is true that the information provided by even a single "extreme member" can be helpful. However, we will consider a different evaluation of the hydrological ensemble predictions, based on the information that could be conveyed by the 90 percentile curve. This is supported by previous publications (Diomede et al., 2008, 2009) where the 90 percentile proved to be a good indicator of the ensemble performance for our basin: based on such studies, for COSMO LEPS coupled with TOPKAPI, the highest quantiles (75-90%) turned out to provide the most informative support to the forecasters in case of high discharge events in the Reno watershed. In the present case, considering the 90 percentile (green) curve, it is evident that only the multi-model is able to reproduce correctly, especially at longer range, the occurrence of two separate and intense peaks.

Based on the local forecasters experience, this would have been a trustable indication of two consecutive intense events. Also, taking into account the strong criticism of Referee #2 on this point, we will try to avoid drawing conclusions based on the analysis of the spread or probabilistic information that would have required an a-priori knowledge of the statistical behaviour of the ensemble systems.

C) Other comments:

Introduction, page 13417, line 27/28: why parentheses for "(and boundary)" ?

The parentheses are due to the fact that this sentence applies both for global and LAM ensembles, and boundary conditions apply only to the latter.

Intro, p.13420, l. 1: "multi-analysis ensemble" It is true that he members use different ICs, but are those really based on different analyses? From the current description of the systems I would assume that the analysis in a sense of assimilation of observations for all LAM members has been in the ECMWF EPS and the LAM EPS do not incorporate different analysis (apart from the effect of mixing two ECMWF EPS start times in the clustering, but thus is more "same analysis approach at different start times"). I have doubts whether "multi-analysis" is appropriate here.

We thank the Referee for having noticed this inaccuracy. Indeed, we used ECMWF EPS members for providing initial and boundary conditions to the mesoscale models. Thus, the global model, the assimilation system and the observations included in the analysis are the same. So it is not appropriate to call it multi-analysis and we will correct this.

Section 2.2: Is there any reference for "Jacobsen and Heise" available?

Section 2.3: Is there any reference for "Noah land-surface model" available?

Since the list of References was already pretty long, we tried to keep the number of citation for the model description as low as possible. But if the Referee believes these two are needed, we will add them.

Figure 6: The graph needs a higher resolution. The annotations explaining the different lines is hardly readable, even when zooming in (it's better in Fig. 7).

We will improve the quality of Fig. 6.

Section 4, p. 13428, l. 17+19 and next page, l. 2+ 12: Should it be "LAMs" instead of "LEPSs"?

LEPSs is correct, since we are pointing out that both the LAM ensembles performed better than the global ensemble. Here we refer to both COSMO LEPS and multi-model as LEPSs (Limited area Ensemble Prediction Systems). Actually, also the multi model is a LEPS.

Section 5 and Fig. 7: It is interesting to note that most members do not have the observed two-peak structure in the discharge prediction. However, if a member shows the two peaks (e.g. P35

members), the effect depends on the LAM, e.g. WRF has the lowest peaks. The authors should include this in the discussion of the multi-model results.

The comment concerning Fig. 7 is supporting our conclusion about the impact of the boundary conditions: at longer range, they are the dominating forcing. Indeed, all the LAM forecasts driven by P35 produce the same two-peak structure. Then, the characteristics of each single LAM produce a "second order" impact, just modulating the peak intensity.

Section 5, p. 13430, l. 4: ". . .it provides a more reliable estimation. . .". This could be interpreted as the EPS being reliable which is a technical term in EPS forecasting. This would need a proof in term of a calculated reliability or another word should be used instead.

Absolutely right. Reliable cannot be used in this context, since we do not have a statistical validation of the ensemble. Anyway, the first sentence of Sec. 5 will be removed (see answer to point B above).

Section 5, p. 13430, l. 15/16. ". . .precipitation patterns remain similar among the forecast driven by the same global representative member.." I think, this statement is too generalizing. E.g: COSMO (m36) member is more similar to the WRF(m3) than to WRF or BOLAM (m36) member, the same with COSMO (m35) and WRF(m23). The stratification along driving members is not that obvious which by the way further supports the use of multi-model approaches. Authors should comment on this in the text

We agree with the Referee. It is not straightforward. We will revise this part in the paper. Differently from shorter forecast range (Fig. 10), there is an evident forcing related to the boundary conditions, but the differences among the models do have an impact too.

Section 5, p. 13431, l. 11-13: I would prefer a less generalizing conclusion about the dominating effect of boundary conditions based on such a case study. The tendency is obvious, but maybe a weaker formulation would be better

Since it is based just on one case study, we cannot draw very robust conclusions. Along this line we will revise the conclusion in order to make clear they are not general.

Section 5, p. 13431, l. 27+28: ". . .have not fully diverged yet. . .,. . .initial perturbations have not grown enough. . .". This is also linked to the general properties of COSMO-LEPS as being based on clustering of IFS-EPS. The latter has its focus on spread in the medium range.

In this section we are discussing the behaviour of the multi-model ensemble and not of COSMO LEPS. Anyway, the point is correct: in the global EPS, perturbations were optimized for the medium-range and the clustering window (between +96 and +120 hour) reflects this goal. This can partially explain why the five forecasts issued by each LAM have not fully diverged yet. Moreover, this is true also for the COSMO LEPS. However, the IC perturbation methodology of the EPS has been recently revised; in particular in 2010 the EDA (Ensemble Data Assimilation) based perturbations have been introduced. The spread/skill relation of the EPS in the short-range has been improved (Buizza et al., 2010). As for the effect of this change on COSMO-LEPS, an evaluation of the spread of the system has highlighted a good performance in the short-range (Montani et al., 2011).

Section 5, p. 13431, l. 29: “close to each other” instead of “close each other”

We will change in close to each other

Section 6, p. 13432, l.12: “multi-analysis”-> see above

OK

Section 6, p. 13422, l. 9 and l 23: “..Reno River basin as an area likely to be affected by. . .”. Possible redundancy or repetition

It will be changed into “…Reno River basin as likely to be affected by…”

Anonymous Referee #2

General comments

This manuscript deals with an interesting subject: the use of mesoscale meteorological ensembles for preparing ensemble hydrological forecasts (EHF). To my knowledge, EHF is far from being largely implemented in operational flood forecasting agencies and one of the reasons is that technological investments needed for implementation are large compared to current performances of EHF (under-dispersion, bias, . . .). Scientific papers that present solutions to increase EHF performances are, in this sense, mostly welcomed.

In my perspective, this paper presents a very interesting set of meteorological modelling tools and a very rich modelling environment. Unfortunately, the analysis of a single flood episode and of a unique watershed does not allow for extracting pertinent and useful knowledge from this rich environment. In my view, the manuscript illustrates that the models are functioning and that results have such or such anecdotic characteristics but there is no solid scientific conclusion that can emerge from a one-site/ one-event methodology. That limits greatly the interest for the manuscript.

A few sentences may better illustrate this point of view. P13416, L25 : " . . . multi-model ensemble provides more informative probabilistic predictions . . . since it characterized by a larger spread . . .". This can be true only if this spread is well calibrated and is associated with the right probability. How can we tell if only a single event is analysed?

How can we do probabilistic forecast and at the same time not analyse the result in a probabilistic framework using probabilistic scores on several events and watersheds (CRPS, ROC, . . .) ? P13427 L10 : "If such a diversity is representative of . . ." : that is exactly the kind of question the paper should try to answer otherwise it is just general opinion not formal science. P13428. "the possible occurrence of high discharge peaks is forecast four or five days ahead (. . .)" How many times it is forecasted but finally did not occurred in other cases ? How can we discriminated which model performs the best if only a case leading to a flood is analysed ? What happens when some models forecast large floods but no hydrological reaction in observed at the end? Which model is the best in this very important practical situation?

Having in mind this major problem in the experimental methodology, my decision is to accept the paper only if major revisions are done. The addition of at least a few events on at least 2 to 3 watersheds seems the only way to produce useful scientific knowledge using this rich modelling environment and must be included in a revised version. A complete rewriting of the discussion/conclusion has to be done based on coming new simulations.

We believe, as addressed by Referee #1, that "*Even though the paper is based on one case study, the results are sufficient to highlight important aspects of the problem and to trigger further research".* Also, we believe we have honestly stressed this point as the main limiting aspect of the present study: since it is based on one case study only, we cannot draw very robust and very general statistical conclusions (page 13434). However, the purpose of the paper is to show that the information conveyed by a multi-model ensemble is helpful and allows to properly address the potential threat associated to the single case study discussed here. Within this context, we tried to analyse in detail the behaviour of the multi-model ensemble. Thus, this paper is supposed to present "just" the starting point of a possible long-lasting research we have planned, whose motivation may need a more precise and clear description (as requested by Referee #1). Our aim is to provide a valuable contribution able to foster further research, indeed. To produce a statistical evaluation of the ensemble performances is surely an important step, but we believe that our multi-model system was not mature for an extensive running period, since we needed first to tackle issues related to the proper design of the system. This work has helped in proposing at least a framework for future investigations.

The starting point of the activity is the availability of an already operational ensemble system (COSMO LEPS), coupled with TOPKAPI for discharge forecasting, whose performance presents room for improvements (as demonstrated by previous statistical analyses - Marsigli et al., 2008), and of several NWP deterministic models, run by different centres that share collaborative activities in the field of hydro-meteorological modelling, as demonstrated by previous publications (Diomede et al., 2008; Davolio et al., 2008). Within this context, our aim is to answer the following question: is it possible to improve the performance of the single-model operational ensemble (COSMO LEPS), also in terms of hydrological predictions, using the information that can be obtained by a multi-model system? To answer this question, the starting point is a comparison between the two ensemble modelling systems, in order to identify possible pro and cons, for a single event. A case study approach does not complete our investigation, but, as we tried to point out in the conclusions, represents just the starting point of a long and complex task.

What the Referee suggests (statistics, calibration) is reasonable from a theoretical point of view, but it seems suitable for a research proposal of several years: models should be implemented operationally and validated both on their daily performance as well as for heavy precipitation/flood events. Calibration of COSMO-LEPS products has recently been operationally implemented, but only after a dedicated study which required few years (Diomede et al., 2010, 2011). The problem of having probabilistic information, probabilistic scores and robust statistics cannot be overtaken by adding just few events or few additional watersheds, as suggested by the referee, but needs a huge effort which is far beyond the scope of this paper, although foreseen in future activities. Moreover, as the Referee says: "*EHF is far from being largely implemented in operational flood forecasting agencies and one of the reasons is that technological investments needed for implementation are large compared to current performances of EHF*". The computational effort is indeed remarkable. Even adding few more events/basins is really demanding and not affordable in few weeks, both for the meteorological models and for the hydrological model that should be re-calibrated in a different watershed.

Thus, taking into account the Referee's comments and recalling our aim, i.e. investigating if and to what extent the multi-model ensemble can provide additional information to the COSMO LEPS for a specific case study, we have deeply revised the paper, limiting the far-fetched conclusions of the previous version. We realized, thanks to the Referee's criticisms, that the terms "ensemble spread", "probabilistic informative predictions", "reliable estimation of uncertainties" and "useful indication for civil protection" are misleading, since they would have required a statistical support that is far for being available. As a consequence, we will remove and rephrase several sentences in order to make clear that we do not know how the ensembles behave in a statistical sense.

We are aware that this limitation will decrease the interest of the manuscript with respect to a more complete paper, but we believe that it is still worth publishing in order to fix a reference for further investigations. Also, we are aware of just few studies dealing with the topic of inter-comparison of different hydro-meteorological ensemble systems, and we believe our effort can represent a useful contribution in the field.

To make the paper more appealing especially from an operational perspective, we added a different ensemble evaluation. Following the Referee's comments, we will try to avoid drawing conclusions based on the analysis of the spread or probabilistic information that would have required an *a-priori* knowledge of the statistical behaviour of the ensemble systems. Instead, we will consider a different evaluation of the hydrological ensemble predictions, based on the information that could be conveyed by the 90 percentile curve. This is supported by previous publications (Diomede et al., 2008, 2009), where the 90 percentile proved to be a good indicator of the ensemble performance: based on a large number of studies dealing with COSMO LEPS coupled with TOPKAPI, the highest quantiles (75-90%) turned out to provide the most informative support to the forecasters of ARPA-SIMC in case of high discharge events in the Reno watershed.

Specific comments

P13417 L6 to L14 : Sentence too long. Please change the text.

We agree and will rephrase the sentence

P13420 L16 : "The poor man's model" is a familiar and wide-spread expression but not pertinent in a scientific paper. Please change the text accordingly.

We disagree with the Referee. Poor-man is widely used in the scientific literature. Just to have some examples of well-known scientists publishing on renowned journals:

Buizza, R; Richardson, DS; Palmer, TN, 2003: Benefits of increased resolution in the ECMWF ensemble system and comparison with poor-man's ensembles, Q.J.Roy.Met.Soc., 129, 1269-1288.

Ebert, EE, 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. Mon.Wea.Rev., 129, 2461–2480.

Anyway, in the present context "poor man" is not completely appropriate since the simulations did not come from operational activity, but were specifically run for this study. So we will remove the term "poor man".

P13426 L18: What does "in excess" mean in this sentence?

In excess mean greater than 20 mm/6h.

P13428 L12: Atypical reference for a well-know interpolation technique.

Right. The reference was included since in that paper the TOPKAPI was implemented using this interpolation technique. We will remove the citation. We believe that this interpolation technique is well-known and does not deserve a specific citation.

P13432 L5 : What is a "best representative member" ?

We agree, it is misleading. We wanted to say: "it is not easy anymore to recognize if a specific representative member of the large scale EPS drives the worst or the best forecast for all the LAMs". We will modify this sentence.

REFERENCES:

Buizza R, Leutbecher M, Isaksen L, Haseler J. 2010. Combined use of EDA- and SV-based perturbations in the EPS. ECMWF Newsletter, 123, 22-28.

Diomede T., C. Marsigli, A. Montani, T. Paccagnella,, 2008: A limited-area ensemble prediction system to drive a flood forecasting chain. Proc. of HydroPredict'2008, International Interdisciplinary Conference on Predictions for Hydrology, Ecology, and Water Resources Management - Using Data and Models to Benefit Society, Prague, Czech Republic, 15-18 September 2008, pp211-214, Edited by Jiří Bruthans, Karel Kovar and Zbyněk Hrkal ISBN 978-80-903635-3-3 publication of the Czech Association of Hydrogeologists, available at http://web.natur.cuni.cz/hydropredict2008

Diomede T., C. Marsigli, A. Montani, T. Paccagnella, 2009: Streamflow ensemble forecast driven by COSMO-LEPS for small-size catchments in northern Italy. Proc. HEPEX Workshop on Post-Processing and Downscaling Atmospheric Forecasts for Hydrologic Applications, Toulouse, France, June 15-18, 2009, p.6 (available online at http://hepex.nmpi.net/files/download/workshops/post-processing/Book_Abstracts.pdf)

Diomede T., C. Marsigli, A. Montani, T. Paccagnella, 2010: Comparison of calibration techniques for a limited-area ensemble precipitation forecast using reforecasts. Proc. of 3rd International Conference on QPE/QPF and Hydrology, 18-22 October 2010, Nanjing, China (available at http://www.wmo.int/pages/prog/arep/wwrp/tmr/QPE_QPF-III.html)

Diomede T., C. Marsigli, A. Montani, T. Paccagnella, 2011: Comparison of calibration techniques for a limited-area ensemble precipitation forecast using reforecasts, Proc. of the Spatial Data Methods for Environmental and Ecological Processes – 2nd Edition, 2011 European Regional Conference of The International Environmetrics Society, Foggia and Gargano (FG), Puglia, Italy, 1-2 September 2011 (available at http://old.unifg.it/spatial/proceedings.asp)

Marsigli C, Montani A and Paccagnella T. 2008. A spatial verification method applied to the evaluation of high-resolution ensemble forecasts. Meteorological Applications, 15, 125-143.

Montani A, Marsigli C and Paccagnella T. 2011b. Recent developments and plans for the COSMO-LEPS system. Proceedings of the 13th ECMWF Workshop on Meteorological Operational Systems, 31/10 - 4/11/2011, ECMWF, Reading, UK, 196-203.