Answer to reviewer # 1

'Unfortunately, this works does not offer new and useful insights, lacking novelty and a rigorous evaluation approach that could eventually turn into valuable information for both the scientific and operational hydro-meteorological communities.'

We believe that the reviewer does not sufficiently justify his/her statement:

- **lacking of novelty**
  To our knowledge there are not yet coupled operational system using regional meteorological ensemble and hydro ensemble in Italy.
  To our knowledge there are approximately only a few Meteorological Centers producing Hydrological operational and pre-operational forecast using ensemble prediction weather forecast in Europe (Cloke and Pappenberger, 2009)
- **rigorous evaluation approach**
  Actually, the aim of the paper is not the statistical evaluation of an operational system but its application to a case study. We believe that a statistical evaluation based on only one case study is not scientifically consistent. However, we agree in using a statistical approach to objectively evaluate the response of the ensemble as we have already done for other cases study (Serafin and Ferretti, 2007, Maiello et al., 2014,. Maiello et al., 2017)

1. **'In my opinion, the novelty of the work cannot rely on the simple combination of downscaled probabilistic meteorological forecasts with hydrological simulations. Moreover, after reading the manuscript it is not clear what's the message the authors want to convey when they generically discuss about the pros and cons of deterministic vs probabilistic forecasting approach. Nowadays the combined use of both is an established practice implemented by many hydro-meteorological centers. That is, the "complementarity" should be substantiated with ad-hoc results and not just advocated. In addition, saying that the probabilistic approach allows for longer forecast periods is absolutely misleading. Finally, in order to put this work in the right perspective I would have referred to the recent and innovative efforts behind the development of WRF-Hydro aiming at fully integrated hydro- meteorological forecasts.'**

   As in the previous statement, we think that the reviewer does not sufficiently justify his/her sentences:
   - **the novelty of the work cannot rely on the simple combination of downscaled probabilistic meteorological forecasts with hydrological simulations**
     We probably missed to clearly specify that this is an off-line coupling of the regional ensemble weather forecast and the hydrological ensemble forecast for the Italian regions. To our knowledge there are not Italian weather forecast centers performing this kind of forecast. Please, let us know which are the hydro-meteorological centers using ensemble weather forecast and ensemble hydrological forecast so that we can refer to them.
   - **Moreover, after reading the manuscript it is not clear what's the message the authors want to convey when they generically discuss about the pros and cons of deterministic vs probabilistic forecasting approach. Nowadays the combined use of both is an established practice implemented by many hydro-meteorological centers.**
     We will clarify this aspect in our conclusions.
   - **That is, the "complementarity" should be substantiated with ad-hoc results and not just advocated.**
     We will add an objective evaluation of the ensemble forecast to support our conclusions.
   - **In addition, saying that the probabilistic approach allows for longer forecast periods is absolutely misleading.**
     We do not agree with sentence. Most of the international weather forecast centers (ECMWF, NCEP etc.) perform ensemble forecast for longer periods than using the high resolution deterministic forecast. Please justify this remark.
   - **Finally, in order to put this work in the right perspective I would have referred to the recent and innovative efforts behind the development of WRF-Hydro aiming at fully integrated hydro-meteorological forecasts.**
     We do agree that the comparison with a well-established and well known (for a long time) hydrological model (WRF-Hydro) fully coupled with the weather forecast is an interesting point to investigate but it is another study. Moreover, we do believe that a different choice regarding the hydrological model

(well referenced too: Tomassetti et al., 2005, Coppola et a., 2007, Verdecchia et al., 2008) does not represent a weakness of the present study.

2. **'The manuscript lacks a quantitative approach in the analysis/interpretation of the modeling results. Several common verification scores should have been implemented in order to assess the performance of the modeling results; mainly precipitation and streamflow. Here I would have also paid special attention on the quantification of the spatial agreement between simulated and observed precipitation (using radar data), which is key for short-term distributed flood forecasting'**
   - **Several common verification scores should have been implemented in order to assess the performance of the modeling results;**
     As already stated at the beginning, we did not perform specific statistical evaluation because the aim of the paper is not the statistical evaluation of an operational system but the application of the system to a case study. We believe that performing a statistical evaluation for a case study only is not scientifically supported. However, a quantitative statistical approach to objectively evaluate the response of the ensemble will be added in the revised version of the manuscript.
   - **Here I would have also paid special attention on the quantification of the spatial agreement between simulated and observed precipitation (using radar data)**
     We will do this comparison. Just to point out, the radar precipitation is a derived product that may have its own error as well. That is why we used rain gauges. However, we will accordingly add the radar dataset and both the products will be considered in the revised manuscript version.

3. **'The content of the manuscript is not well organized (i.e., section order). In general, you should present first data and methods (i.e., numerical models), define the skill metrics and/or indices, and at the end you interpret/discuss the results. The language should be improved. This is not just a matter of typos, grammar mistakes, and unclear sentences. Finally, many figures (e.g., Fig. 3, Fig. 7, Fig. 11) and Table 1 are really not necessary. Please note also that the geographical location of the two Italian regions (i.e., Umbria and Abruzzo) is not shown in any map. The same apply for the hydrometric stations. '**

   - The content of the manuscript is not well organized (i.e., section order).
   The organization of the paper is basically what the Reviewer#1 is suggesting. We first presented the case study and the data. Then, it follows the methods used i.e. the models.
   We decided to separate the presentation of the models and their results. WRF and its results presented first and then it follows CHyM with related results.
   Therefore, the paper organization is the following:
   1. Introduction
   2. Case study, that is the data used for this study, as you are suggesting
   3. WRF Numerical model, that is the method used for the weather forecast
   -Ensemble weather forecast results
   4. CHyM Hydrological model, that is the method used for the hydrological forecast
   - Method used to evaluate the hydrological results: BDD index
   5. Hydrological model results using different forcing: ensemble mean, all members and deterministic
   6. Conclusions

   - **The language should be improved.**
     We will carefully review the whole paper paying particular attention to the grammar and the language.
     - Finally, many figures (e.g., Fig. 3, Fig. 7, Fig. 11) and Table 1 are really not necessary. Please note also that the geographical location of the two Italian regions (i.e., Umbria and Abruzzo) is not shown in any map. The same apply for the hydrometric stations.'
     We agree with you, Figs. 3, 7 and 11 will be removed from the paper. We believe that table 1 helps understanding the performed experiments. We will add the location of the Italian regions Umbria and Abruzzo.

4. **'I have some remarks on the model setup and configuration: - The use of 1° GFS forecast is not fully justified in my opinion because other deterministic and probabilistic products (e.g., ECMWF) are available at higher resolution. - Domain definition, grid resolution, and physical parameterizations have a large influence on model results. Did the authors made preliminary tests to check their impacts on the selected events? This is key for a sound modeling strategy, especially from an operational perspective. - The definition of the benchmark configuration is not clearly discussed. If I look at Fig. 1 of Pichelli et al., 2017, the 1km domain (D3) does not fully cover the study area of this work, am I wrong? In the same paper it is mentioned that the operational setup of WRF-CETEMPS is different from the one shown in Fig.1. Further, it seems that the high-resolution setup uses GFS later boundary forcing with different resolutions (i.e., 0.25° instead of 1° resolution) and different physical model parameterizations. Finally, it seems not completely justified to directly nest the 9km WRF into the 1° GFS forecast. I would have expected an intermediate step to reduce lateral boundary effects. In general, these aspects of the work are not clearly explained.**

   - **'I have some remarks on the model setup and configuration: - The use of 1° GFS forecast is not fully justified in my opinion because other deterministic and probabilistic products (e.g., ECMWF) are available at higher resolution. - Domain definition, grid resolution, and physical parameterizations have a large influence on model results. Did the authors made preliminary tests to check their impacts on the selected events?**
     Based on our long experience in using several numerical models (MM4, MM5, WRF, Harmonie), we run operational weather forecast since 1998 (Paolucci et al., 1999) and several paper published on this topic, we defined the model set up and performed several experiments using different ICs (ECMWF and GFS) and different parameterizations. The best results we end up with is the one presented in this paper.
     Unfortunately, the NCEP archive allows for retrieving only 1° forecast and analysis, this why we used the 1° GFS. For what concerns ECMWF we performed several tests using the 50 members of ECMWF ensemble at 0.125 but the results obtained were not satisfactory.
   - **'The definition of the benchmark configuration is not clearly discussed. If I look at Fig. 1 of Pichelli et al., 2017, the 1km domain (D3) does not fully cover the study area of this work, am I wrong?**
     The Pichelli's work is on the Pò Valley that is on a Valley delimited by the Mountains (Alps and Apennine) that is complex orography region as Abruzzo region is. Therefore, based on the Pichelli's work we set up the deterministic operational forecast at 1km (http://magritte.aquila.infn.it/meteo/ecmwrf-2way/) over Abruzzo region. We used this operational deterministic configuration to run an 'ad hoc' deterministic forecast using 0.25 GFS, ICs and BCs, for this event and we used it as benchmark.
   - **In the same paper it is mentioned that the operational setup of WRF-CETEMPS is different from the one shown in Fig.1.**
     Please see the previous answer.
   - **Further, it seems that the high-resolution setup uses GFS later boundary forcing with different resolutions (i.e., 0.25° instead of 1° resolution) and different physical model parameterizations.'**
     Yes, we used the best deterministic forecast produced by GFS at 0.25° and the best available GFS ensemble forecast, that is at 1°. Since we used the deterministic high-resolution forecast as benchmark we decided to do not downgrade it. The use of different parameterizations, as you just stated, is driven by the different resolution. We have to use a cumulus convection parameterization at 9km which is not necessary at 1km because convection is explicitly resolved at this resolution.
   - **'I would have expected an intermediate step to reduce lateral boundary effects. In general, these aspects of the work are not clearly explained.'**
     We agree with you. Generally, it is better to use an intermediate domain, if going down from 1° to 9km, for providing BCs to the nested domain. We performed several simulations but there was not any improvement in the results with respect to the direct nesting into the GFS ICs. Therefore, based on these results we decided to perform the ensemble forecast directly nesting the 9km to the 1° GFS ICs. We will add an explanation about this in the reviewed version of the paper.

5. **‘I have several remarks concerning the adopted discharge index and the discussion of the related results: - Authors consider the definition of the BDD index necessary due to the lack of discharge measurements. This contradicts the definition (Eq. 2) of the index itself, which is based on discharge values! - What’s the equation used to calculate the hydraulic radius as a function of the drainage area? The comparison between Fig. 10 and Fig. 9 is not intuitive. - I do not fully agree with the interpretation of Fig. 10. I see a good agreement in the timing even for those stations heavily impacted from hydropower production (i.e., Vomano and Todino). I also think that the mismatch for Pescara River could be due to some error in the observed atmospheric forcing at the local stations. That’s why a more careful evaluation of the atmospheric forcing would have provided more useful insights.**

   - **Authors consider the definition of the BDD index necessary due to the lack of discharge measurements. This contradicts the definition (Eq. 2) of the index itself, which is based on discharge values! - What’s the equation used to calculate the hydraulic radius as a function of the drainage area?**

     The observation is very appropriate and we thank the referee for this; we tried to summarize in one sentence two different problems, leading to a lot of confusion. The first problem deals with the difficulties to calibrate the discharge predicted by any hydrological model with observed data. Discharge observations in continuous time series are often missing, especially for small basins.

     A different problem is to use the predicted discharge for flood alert mapping, as it is not straightforward to establish a threshold level above which a critical event is to be expected; in addition such threshold level should be calculated for each grid point because it depends on the size of the river bed in the selected point. To overcome this second problem we tested different general definition of an alarm index and, after simulating different case studies occurring in different basins of different size, we find that a suitable definition could be the ratio between the maximum value of the predicted discharge within a given time interval and the square of hydraulic radius that is a “measure” of the river cross section for the selected point. The definition of BDD index has also a simple physical interpretation: it represents the average precipitation (more specifically the precipitation available for the runoff) drained by each grid element from the upstream basin.

     The BDD index is based on Eq. 2: in this equation, the used discharge value is not the measured value, but the discharge computed by the CHyM model, forced with observed raingauges data as input.

     As for many other models (for a general reference see Singh and Frevert, 2002) the hydraulic radius can be approximated as a linear function of drained area. In particular $R=\beta+\gamma D^{\delta}$ *where* β, γ *and* δ *are empirically established and the value of* δ *is very close to 1. If the area is measured in Km2, typical values taken from literature are* $\beta=0.0015$ *and* $\gamma=0.05$*, while (for a general reference see Singh and Frevert, 2002).*

   - **The comparison between Fig. 10 and Fig. 9 is not intuitive. - I do not fully agree with the interpretation of Fig. 10. I see a good agreement in the timing even for those stations heavily impacted from hydropower production (i.e., Vomano and Todino).**

     We will explain in details the two figures. For what concerns fig.9 the four red triangle-shaped, thin-bounded signs indicates the relevant hydrometers where the red hydrometric threshold has been exceeded; in particular, among the involved rivers, there are Vomano, Tordino, Saline and Pescara. In figure 10, the normalized water level and BDD time series along the aforementioned rivers, for different hydrometric station grid-points, are shown.

     Generally, hydroelectric power installations can heavily impact the flood dynamics along a river basin, but the key parameters to be considered are various , such as the relative importance of the drained areas, the water storage capacity and the position of the reservoirs within the basin. Nevertheless, the effect highly depends on the initial reservoir filling rate, which is unknown. If the reservoirs are already full before a flood, no (gated spillway) or limited (ungated spillway) flood routing is possible (Jordan et al.,2012). Unfortunately, in Abruzzo region we are not aware of how the hydroelectric systems are managed. In this particular case, the hydroelectric power plants of Provvidenza and Piaganini are located upstream (Figs. 1 and 2, below this section), respect to the areas involved in the event, where also precipitation maxima occurred. Probably, in this case, the effect of the hydroelectric power plants is *negligible and this sentence is confirmed by the good agreement in the timing shown by Fig.10.*

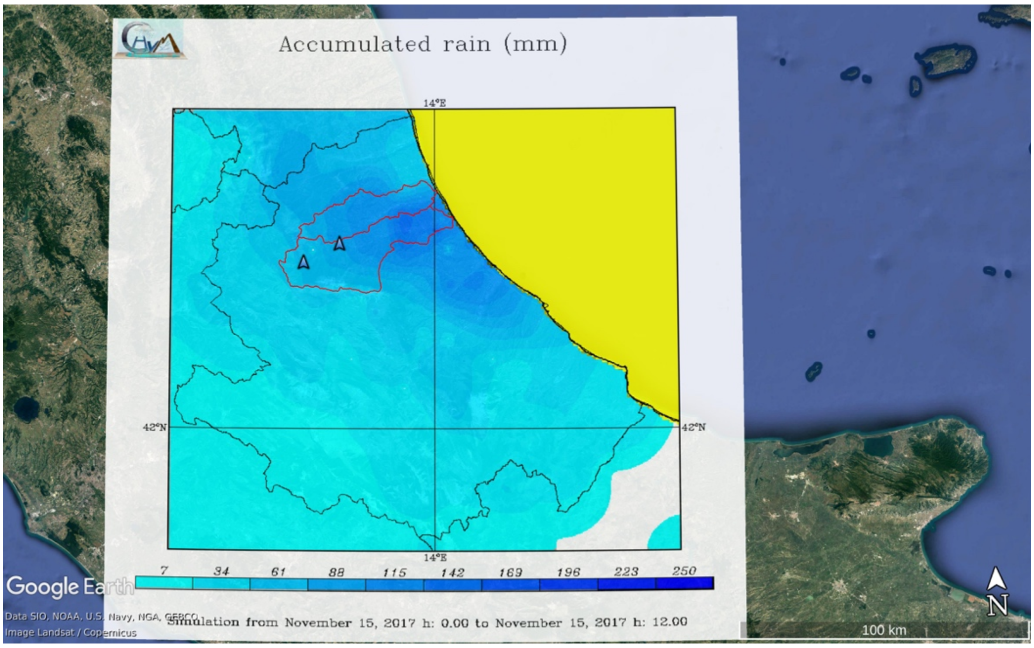

Figure 1: The map represents the accumulated precipitation from 0 UTC to 12 UTC on 15th Nov 2017, as measured by the raingauges network and spatialized over the region by using the Cellular Automata-based techniques. The area enclosed in the red line is the boundary of the Tordino basin. Blue triangles indicate the position or the Provvidenza and Piaganini dams.

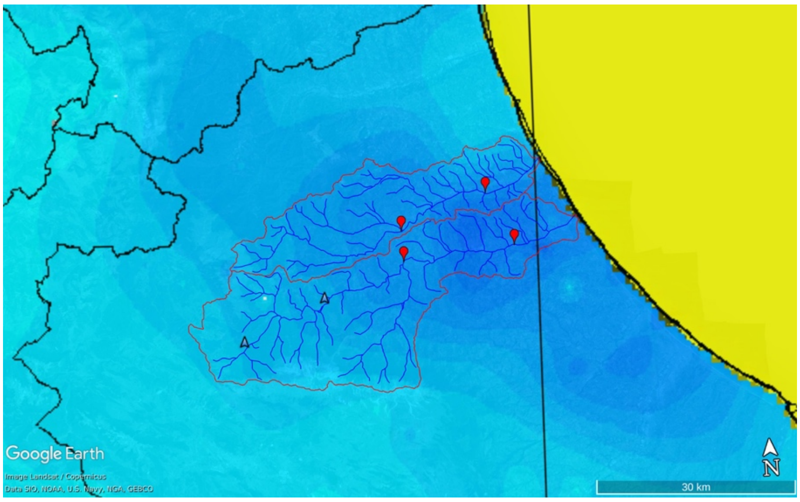

Figure 2 is a zoom of the figure 1, where the Tordino drainage network is indicated by the blue lines. Together with the dams, relevant hydrometric station are also indicated through red pinpoints.

- **I also think that the mismatch for Pescara River could be due to some error in the observed atmospheric forcing at the local stations. That's why a more careful evaluation of the atmospheric forcing would have provided more useful insights.**
  Please clarify what you mean by 'error in the observed atmospheric forcing'. Are you referring to the data quality at Pescara station?

6. **For instance, authors interpret the results saying that the mismatch between "observed" and "simulated" BDD index is due to precipitation occurring only on a very small area and not capture by the model. What do you mean with "small"? The high-resolution simulations are at 1km! I suspect that you can get the same issue if you go down to 100m resolution. Again, if you do not carefully evaluate the atmospheric simulations it is difficult to provide convincing interpretation of**

the BDD index. Finally, I would also remark that authors talk about "overestimation" and "underestimation" of the BDD index using as a reference the model results driven with observed (interpolated?) precipitation. I am fine with this as long as you cross-validate local precipitation measurements with other sources of information, e.g., spatially distributed information obtained from radar retrievals.

- **For instance, authors interpret the results saying that the mismatch between "observed" and "simulated" BDD index is due to precipitation occurring only on a very small area and not capture by the model. What do you mean with "small"?**
  The Calvano river is a very small basin (35 km^2) and is located close to Vomano final segment, southward. The distance between the two rivers is about 2.5 km in the upper part of the Calvano's path and almost 6 km in the two mouths. Being so close, the rain spatial distribution plays a very important role: an error of even 1 km can significantly affect the forecast. Nevertheless, the Civil Protection early warning system is referred to "warning areas", rather than the single river segment or the single catchment area. For this reason, in a Decision Support System perspective, is important to assign the correct alarm state at warning area level, rather than meticulously focusing on the single catchment.
- **Again, if you do not carefully evaluate the atmospheric simulations it is difficult to provide convincing interpretation of the BDD index.**
  Please clarify this sentence. What do you mean by carefully evaluate atmospheric simulation? To objectively evaluate the weather forecast, as we already said, we will use skill statistical metrics, is this what you are suggesting?
- **I am fine with this as long as you cross-validate local precipitation measurements with other sources of information, e.g., spatially distributed information obtained from radar retrievals.**
  We will compare the results with the retrieved radar precipitation, but again being the radar precipitation a retrieved product it is affected by error as much as other observed parameters.

7. **'I would expect the same kind of curve when I look at the black ("observed") lines in Fig. 14 and the red ones in Fig. 10, am I wrong somewhere? One of the main conclusions is that the uncertainty in the BDD index is underestimated if you do not perturb the parameters of the hydrological model. This is intuitive and this is the reason why you should take both ("atmospheric" and "hydrologic") into account. In my opinion this opportunity was missed in this work.**
   - **I would expect the same kind of curve when I look at the black ("observed") lines in Fig. 14 and the red ones in Fig. 10, am I wrong somewhere?**
     The curves appear different because of the different temporal scale. Moreover, figure 10 shows the normalized index values, whereas fig.14 shows values in mm/h.
   - **One of the main conclusions is that the uncertainty in the BDD index is underestimated if you do not perturb the parameters of the hydrological model. This is intuitive and this is the reason why you should take both ("atmospheric" and "hydrologic") into account. In my opinion this opportunity was missed in this work.**
     Based on Cloke and Pappenberger (2009) this is not an intuitive conclusion. The lack of hydrological ensemble forecast does not allow to make such statement. If you are aware of different published conclusions please let us know.

References

Jordan F.M., Boillat J.-L. and Schleiss A. J., Optimization of the flood protection effect of a hydropower multi-reservoir system, Intl. J. River Basin Management Vol. 10, No. 1, pp. 65 – 72, 2012

Maiello I., R. Ferretti, S. Gentile, M. Montopoli, E. Picciotti, F. S. Marzano, and C. Faccani: Impact of radar data assimilation for the simulation of a heavy rainfall case in central Italy using WRF–3DVAR. Atmos. Meas. Tech., 7, 2919–2935, 2014 www.atmos-meas-tech.net/7/2919/2014/ doi:10.5194/amt-7-2919-2014

Maiello I., S. Gentile, R. Ferretti, L. Baldini, N. Roberto, E. Picciotti, P. P. Alberoni, and F. S. Marzano: Impact of multiple radar reflectivity data assimilation on the numerical simulation of a flash flood event during
the HyMeX campaign. Hydrol. Earth Syst. Sci., 21, 5459–5476, 2017 https://doi.org/10.5194/hess-21-5459-2017
Searfin S. and R. Ferretti,: Sensitivity of a Mesoscale Model to Microphysical Parameterizations in the MAP SOP Events IOP2b and IOP8. JAMC, 46, 1438-1454, 2007. DOI: 10.1175/JAM2545.1
Singh VP, Frevert DK (2002) Mathematical models of smallwatershed hydrology and application. Water ResourcePublications, LLC, Highlands Ranch, Colorado, USA