Ensemble flood forecasting considering dominant runoff processes: II. Benchmark against a state-of-the-art model-chain (Verzasca, Switzerland)"

Authors replies to RC2:

We want to thank the reviewer for his/her assessment of our manuscript. In the following we give our answers to the comments and recommendations that have been raised. Reviewer comments RC are **bold,** our reply AR is in *italic*. Insertions in the revised manuscript MI are underlined.

GENERAL ASSESSMENT

RC: … the results are based on a very limited dataset (just one summer season for the meteorological analysis and about twenty events for the hydrological analysis), that may result as not sufficient to support the interpretations and the conclusions.

AR: This issue has been raised also by the reviewer 1 and by the reviewers of the companion paper by Antonetti et al. (2018). We are of course aware, that more basins and longer periods of evaluation are always welcome. NHESS (but also HESS) is in this respect a journal that regularly publishes case studies (e.g. Kobayashi et al., 2016; Cane et al., 2013), preliminary assessments (Picciotti et al., 2013) or intercomparison of approaches during limited period of time (e.g. Davoli et al., 2018; Li et al., 2018). Having targeted NHESS as journal for disseminating our experience, this study is designed to benchmark a state-of-the-art and operational calibrated hydrological prediction system(PREVAH-HRU) against a newly developed event-based system that is configured without calibration requirements (RGM-PRO). This has been evaluated during a representative flood season and in case a nested basins where PREVAH-HRU simulations have been running and collected in real-time (so no tailored experiment here, 100% data from a deployed system), while RGM-PRO simulations have been completed as reforecast experiment in the framework of a master project (August 2016 to February 2017) by the lead author. With this approach we can learn about the quality of the novel approaches from different perspectives at the same time. The transfer of experience to another catchment and climatic region is presented in the companion paper by Antonetti et al. (2018), while in prior studies we shown how PREVAH-HRU behave in case of long-term analyses (e.g. Addor et al., 2011 and Zappa et al., 2011). This is al why we take so much time and space in order to discuss these findings with respect to our previous studies. Furthermore, As far as the length of the investigation period is concerned, some limitations arise from the use of COSMO-E and COSMO-1. MeteoSwiss decommissioned after several years the antecedent operational NWP COSMO-2 and COSMO-LEPS in 2016. As we want to make our systems operational, it was for us important to focus on a first analysis with the new NWPS that we receive and archive in real-time since February 2016.

In the revised manuscript we will better declare our choices concerning selection of basins and investigation period. Furthermore we will try to indicate to which extent the available season is representative with respect to log-term discharge observations in the target area. We have currently unfortunately no capacity to extend the analyses beyond 2016, and in case of 2018, this would not beneficial since no severe flood event occurred.

GENERAL COMMENTS

RC: (1) It is not so clear the main goal of the manuscript, with respect to the companion paper. On the one hand, if the focus of the paper is on the meteorological input (as stated by authors), therefore the dataset may results as quite limited to test the improvements of the new meteorological forecasting tools (why have authors not considered a larger period of data availability of COSMO-1 and COSMO-E? For instance, the years 2016-2018?). Actually, if the focus of the paper is the evaluation of the performance provided by the new meteorological chains, then the computation of the statistical scores could be carried out over a longer period, without the need to perform a comparison with the older forecasting chains (model benchmarking), given that the statistical scores are able to give an objective evaluation of performance for the tested meteorological forecasting tools. On the other hand, if the focus of the research is on the performance of the new meteo-hydrological chain, therefore the dataset seems to be not so significance in terms of flood events, in particular with respect to the operational aims of civil protection authorities (just one event with a 2-yr return period in the investigated dataset). The limited amount of data here used for the statistical analysis may not justify a separate manuscript with respect to the companion paper. Maybe, the investigations and results shown in the present manuscript could be synthesized (as done, for instance, in Section 6.6) and added to the companion paper in order to enlarge the statistics for the proposed coupling of RGM-PRO approach to COSMO-1 and COSMO-E.

AR: We thank the reviewer for this (meaningful) remark. Our original manuscript was probably not enough clear in this respect, but both companion papers focus in first order on the hydrological predictions chains evaluated. The focus on the meteorological input we declares, refers on the focus in the presentation of the data and method used As we are willing to avoid way too large overlaps between the two paper, we decided to focus the introduction of methods of paper I by Antonetti et al. on the setup of RGM-PRO and on Combiprecip and on classic measures of agreement in hydrological modelling, while in this manuscript by Horat et al. we emphasize the method section on numerical weather prediction (NWP, including a small assessment) and on the common metrics used for verification of (hydrological) ensemble predictions . We are sorry that the reviewer expected a "more meteorological" contribution and we will make our best to make our choice fully intelligible in the revised manuscript.

We kind of agree with the reviewer suggestion to make a synthesis of this manuscript and include it in the companion paper. We explored this option, but we have never been happy with it. As already stated, this would have implied to include the details on PREVAH-HRU, on the NWPs and on the verification metrics in the companion paper, making it way to long (and coincident with the 106 pages thesis by the lead author of this manuscript). If the editor is the opinion that we should explore a merging of the manuscripts, then we can work in this direction.

RC: (2) The meteo-hydrological model coupling is used as a verification tool for the new meteorological model chains at higher resolution. But, by the hydrological perspective, it seems that there are not enough high-impact events in the investigated period. Moreover, it could be questionable that the false alarms are "realistically" evaluated (Pag.8, lines 7-9), given that the investigated period is the summer season. I mean, which is the soil saturation of the study area in summer? Is summer a dry season for the study area or the soil saturation is quite high in summer so that a light/moderate rain event could trigger a flood event? Or are there floods in summer only due to extreme rainfall events which cause rapid surface runoff without rainfall infiltration? Authors should add "hydrological" details about the occurrence of flood events in summer for the selected catchments.

AR: We will add the required details on flood generation in the Verzasca area in summer. As the basin is in a region with steep topography the soils are quite shallow and, as discussed in Zappa et al. (2011) initial conditions are not very sensitive with respect to controlling peak discharge in summer.

RC: (3) The introduction (i.e., Section 1) could be shortened (for instance, lines 14-33 at Pag.2; lines 13-34 at Pag.3; lines 1-31 at Pag.4). Some issues are repeatedly discussed and too much detailed descriptions of past studies are provided, even though not strictly related to the contents and methodologies proposed in the manuscript. Thus, a synthesis may result advantageous. Moreover, the contents may appear as dispersive (too general) with respect to the context of the study area and the proposed forecasting methodologies. The contents of Section 1 should focus on contents which show similar features to the present study. The citation of past studies should highlight the feasibility of those approaches with respect to spatial and temporal characteristics of phenomena (for instance, catchment dimension, return time of the basin, forecast lead time), focusing on the similarities with the present manuscript. In the current form, this section seems as a general review of the flood forecasting subject.

AR: Both reviewer raise this issue. A more concise and specific introduction will be prepared.

RC: (4) I guess that the hourly runoff climatology of the period May-August 2016 was used as reference climatology to carry out the statistical analyses (for instance, to compute the quantiles of Figures 6-8). Is the May-August 2016 runoff climatology statistically meaningful with respect to a longer climatology (for instance, some decades) for the selected study area? Section 5 provides a very detailed analysis of the performance

for the tested forecasting chains. Nevertheless, it is not so evident that these performances are significantly different. Even, some scores provide outcomes in contrast to the companion paper (for instance, the process-based forecasts are not better for the nested sub-catchment). The limited dataset may hamper a solid comparison.

AR: Yes, the reviewer is correct in his assumption concerning the used "climatology". This choice and limitation due to the shot duration of the data set affects in equal ways both the RGG-PRO and the operational PREVAH-HRU chain. We are of the opinion, that even if a sound indication on the absolute quality of both chain cannot be provides, we can provide very clear indications on the difference in quality between the RGM-PRO and the benchmark (which has already been verified for longer time series using the predecessors of COSMOE and COSMO1 as forcing). For us is therefore a quite interesting and far reaching finding, that RGM-PRO can "compete" and/or "keep the pace" with a state-of-the-art system, even if no calibration is needed. We find this a very good news after years of theoretical advances on prediction of flash-floods in ungauged areas, to show that a "PUB" inspired approach yields similar quality in terms of forecasting skill in real-time mode when compared to a calibration oriented approach. Of course we can still only speculate on the results for long-term operations of such systems. Unfortunately reforecasts of such NWP are (computationally) expensive and cannot be provided.

RC: (5) Sections from 6.3 to 6.6 go into a detailed analysis about comparisons of the proposed new forecasting chains with previous studies. However, models, data input, study areas and investigated period are not always the same. Therefore, the content of these sections may result too long, redundant and not so interesting with respect to evaluation of the proposed forecasting chains. The discussion recalls results and trends which are general and well known in the past specific literature (in particular Section 6.4). Authors should highlight the original contribution of the proposed forecasting chains and summarize the comments on the comparisons (for instance, authors could move each specific comments of Section 6.5 in a position within the manuscript where that issue has already been discussed, rather than devote a specific section to comment all the issues of the comparisons).

AR: Thanks for this comment and useful suggestion that we will implement in the revised version.

SPECIFC COMMENTS

Pag.1, Lines 20-22: The comment on the performance of the proposed model chains should stress the feasibility of these chains with respect to the dimension of the study area. This point of view for the discussion of results could be an added value for the present manuscript.

AR: We will consider this suggestion in the revised manuscript, thanks!

Pag.2, Lines 1-2: This statement should be based on a larger dataset.

AR: The sentence now reads: "The findings of the two studies as obtained from a set of data limited to one flood season…"

Pag.2, Line 11: The meteorological perspective is not so deeply investigated in the manuscript.

AR: We will arrange this as replied to you general comment (1).

Pag.3: Line 2: Authors should specify the reasons for the suitability just for catchments with areas up to 1000-2000 km2.

AR: The reasons are: a) increase of chance of having more disturbed catchments. b) delineation of process maps for such large areas. c) larger catchments are not prone on the kind of flash-floods triggered by local thunderstorms, that we are focussing on. These three points will be added in the revised manuscript.

Pag.3, Line 18:Authors should specify the country of FOEN.

AR: Swiss Federal Office for the Environment (FOEN)

Pag.3, Line 29: Authors should specify the year of the event.

AR: 2007.

Pag.6, Line 10:Authors should specify the meaning of the acronym "WSL".

AR: Swiss Federal Institute for Forest, Snow and Landscape Research

Pag.6, Line 10:Authors should specify that PREVAH is a semi-distributed hydrological model (as done in the abstract).

AR: Done.

Pag.7, Line 9: Should "and Avalanche Research SLF." be "and Avalanche Research (SLF)."?

AR: Done.

Pag.7, Lines 15-22: Is the configuration of the COSMO model changed with the in-crease of horizontal resolution (from 2.2 to 1.1 km)? Authors should add details and references about this issue.

AR: The "Numerical weather predictions" section will be re formulated as follows. The two new tables will be added as supplementary material.

MeteoSwiss developed a configuration of the COSMO model (Steppeler et al. 1998) with 1.1 km grid spacing, the COSMO-1 (Fuhrer et al., 2014). It runs as deterministic model and is initialised from its own assimilation cycle using the nudging scheme. Forecasts are calculated

every three hours in a rapid update cycle with a forecast range of 33 hours and once per day (03 UTC forecast) out to 45 hours. This setting was operationalised in spring 2016 and replaced the former COSMO-2 with 2.2 km grid spacing and a configuration similar to that described by Baldauf et al., 2011. Configuration changes from COSMO-2 to COSMO-1 are listed in Table S1. As its predecessor, COSMO-1 assimilates radar-derived QPE using latent head nudging. Latent heat nudging is able to considerably increase the accuracy of the precipitation forecast during the first 6 to 12 hours of the forecast. The boundary conditions are taken from the newest available ECMWF (European Centre for Medium-Range Weather Forecasts) high resolution forecast (HRES).

Table S1: Configuration of the deterministic models COSMO-2 and COSMO-1

Configuration	*COSMO-2*	*COSMO-1*
Grid spacing	*2.2 km*	*1.1 km*
Levels	*60*	*80*
Convection parameterization	*Shallow convection*	*None*
External parameter fields from	*GLOBE*	*ASTER*
Num. diffusion for wind	*On*	*Off*
Boundary conditions	*COSMO-7*	*IFS HRES*

In addition to the deterministic COSMO-1, the ensemble system COSMO-E with 2.2 km grid spacing was operationalised in May 2016 (Klasa et al., 2018). It is initialised twice per day and has a lead time of 120 hours. The assimilation cycle uses an ensemble transform Kalman filter approach (Schraff et al., 2016). The boundary conditions are taken from randomly selected 20 members of the ECMWF ensemble forecast (ENS). It uses the SPPT scheme to simulate the effect of the model uncertainty. At MeteoSwiss, COSMO-E replaces COSMO-LEPS (Marsigli et al., 2005; Montani et al., 2011), which has a lower resolution with 7 km grid spacing. These and further configuration changes are listed in Table S2.

Table S2 Configuration of the ensemble prediction models COSMO-LEPS and COSMO-E

Configuration	COSMO-LEPS	COSMO-E
Grid spacing	7.0 km	2.2 km
Levels	40	60
Ensemble scheme	Parameter perturbation (PP)	Stochastically perturbed

		physics tendencies (SPPT)
Convection parameterization	Tiedkte	Shallow convection only
Subgrid scale orography	On	Off
Latent head nudging	Off	On (first 2 h)
Initial conditions	Interpolated from IFS ENS	Nudging analysis cycle
Boundary conditions	IFS ENS	IFS ENS

Pag.7, Lines 23-27: Is the confguration of the COSMO-based ensemble changed with the increase o fhorizontal resolution(from 7km of COSMO-LEPS to2.2 kmof COSMO-E)? Authors should add details and references about this issue.

AR: see rely to the previous point.

Pag.8, Lines 1-3: This sentence is not clear.

AR: We will re-arrange it to improve our message.

Pag.8, Lines 6-7: Does a threshold exist to identify major food events (namely, food events which are of interest for the authority in charge of the public safety)? How many major food events occurred in summer 2016 for the study area?

AR: Such thresholds are published by the SWISS FOEN (https://www.hydrodaten.admin.ch/en/2605.html). In 2016 there was one observed value il danger level 2 (450 m3/s). In 2018 the highest discharge until October 3rd was 157 m3/s, in April). In 2017 the biggest event was 2017 m3/s. All in all a quite flood poor three year span in that region., with 2016 being the most flashy season.

Pag.8, Lines 23-24: The comparison may results as not fully proper, given that the observed rainfall input is different for the two chains. Why has the same input not been used for both the chains?

AR: The PREVAH-HRU chain is using the same input rainfall from observations in real-time since 2007, and for this reasons we use it as benchmark. RGM-PRO has been designed to work with the most advanced rainfall product of MeteoSwiss (Combiprecip). An assessment on the difference between the two rainfall inputs is presented in Andres et al. (2016, referenced in the manuscript). We will add a sentence to comment on this.

Pag.9, Line 13: Is the hourly runoff climatology of the period May-August 2016 statistically meaningful with respect to a longer climatology (for instance, some decades) for the selected study area? It could be useful to show a comparison between the reference climatology of this study and a historical ones for the selected catchment.

AR: Good point. We will present this analysis in the revised manuscript .FOEN published such statistics online (https://www.hydrodaten.admin.ch/lhg/sdi/jahrestabellen/2605Q_16.pdf)

Pag.9, Lines 14-16: Are the statistical scores computed at each hourly time step of the simulation event? I mean, is the threshold exceedance evaluated each hour and the corresponding score computed at an hourly time step, then averaged over each lead time window? Or is the threshold exceedance evaluated just one time within the whole lead time window? Please clarify.

AR: Each hour of each forecast is evaluated by itself and its lead time since start of the forecasts is tracked as attribute. Integral scores are later averaged for all forecasted hours with identical lead time. This explanation will be added to the section.

Pag.10, Lines 10-13: The description of the Brier Score decomposition could be omitted, given that it is not discussed in the main manuscript.

AR: We decided to keep the definition of the components in the manuscript, as the results are presented in the supplementary material.

Pag.11, Lines 20-21: Which is the spatial domain (Switzerland? Verzasca catchment?) over which the scores shown in Fig.3 were computed? Please specify. The scores for POD and FAR are not so satisfying, especially for the higher thresholds (namely, rainfall events which likely trigger food peaks). The FAR scores may results quite high with respect to the usefulness for operational decisions of civil protection authorities. Could authors add a comment to justify this results? Which is the rainfall threshold that trigger major flood events for the study catchments?

AR: MeteoSwiss had a very short phase where both systems could be operated and maintained. These are the numbers averaged over Switzerland that have been obtained. In our purpose, this should show the continuity that the new models offer with respect to the predecessors. Floods are triggered by rather intensive rainfall events (above 20 mm per hour). Rainfall based thresholds are currently not used for flash-flood warning in Switzerland.

Pag.12, Lines1-2:Why are the BSS values not shown in Fig.4 (or discussed) for the thresholds higher than 10 mm (as done in Fig.3)?

AR: The presented rainfall statistic originates from the standard output of the diagnostic and verification routines by MeteoSwiss, where 10 mm/h is the highest threshold delivered for the Brier Skill Score.

Pag.12, Lines 14-15: Please specify that the cited scores refer to runoff data and the quantiles refer to the hourly runoff climatology of summer 2016 (in case of my interpretation is right).

AR: Will be done.

Pag.13, Line 14: The panels "b" and "c" of Fig.8 are very friendly to convey the best performing method, but this visualization does not allow to evaluate if the difference of performance is significant in term of ROCa.

AR: You are right. These panels integrates Figure S1 of the supplementary material, where you can better see how much one of the systems outperforms the other one. Figure 8 is specifically designed to answer the question on the best performing method. The link between S1 and Figure 8 will be strengthened and we will elaborate in this respect on the issue of significance.

Pag.13, Lines 16-19: Authors should try to justify this result. May the reason for this result be ascribed to the larger dimension of the Verzasca catchment (with respect to its sub-catchment), which can allow to "average" spatial errors in the localization of the rainfall feld?

AR: Thanks for the remark. We will elaborate on this.

Pag.13, Lines 21-22: The magnitude of the event (i.e., 2-yr return period) should also be cited here (as done at Pag.14, Line 4). Authors could add a comment about the frequency and amount of the flood peaks corresponding to more intense events for the selected catchments.

AR: We agree, as already commented. We add here, that is already "good" to have a 2-years event in the selected years, as the two years afterwards no such event occurred.

Pag.13, Lines 24-30: The significance of the comparison may result as weak, given that it is discussed for two forecasting chains that differ for the model and input used. Authors should add a comment in order to justify this result with respect to model characteristics and data input.

AR: This follows a previous comment on the rainfall input. The reply to the previous point will be extended also here.

Pag.15, Line 6: With this statement authors recognize that the discussed results and conclusions may be invalidated by the limited dataset that has been used to test the proposed new forecasting chain. Actually, the usefulness and added value of the new forecasting chain are questionable due to the limited test period. The use of a different dataset could provide different (and opposing) results.

AR: We think that "invalidated" not the right term here. As stated before, the chains are different, one of do not rely on calibration, and, during the same period and same constraints, similar skill is found. For us this is an advance with respect to other published approaches for ungauged areas, that have been never benchmarked against state-of-the-art chains. Other authors have been working for years on single extr4me events that have been re-forecasted with and without numerical models, our new approach is quasi-operational.

We dared it, we get a promising result, we acknowledge that a short period is a limiting factor, but we think this is a useful communication.

Pag.16, Lines 1-19: The detailed comment on the comparison results does not lead to a clear conclusion about which chain should be preferable. Some scores provide opposing outcomes (for instance, in contrast to the companion paper, the process-based forecasts are not better in the sub-catchment), then this analysis may result as inconclusive.

AR: See previous comment. We are not looking the best method, best would be to have overall the possibility of using calibrated model. We show that the RGM-PRO approach can compete with a calibrated model in real-time mode. This is for us enough conclusive as compared to the current literature.

Pag.19, Line 25:Authors should specify the magnitude of these 22events with respect to the catchment climatology.

AR: See previous replies

Pag.20, Lines 1-4: This result is quite general and recalls several past studies. Here, it appears as a local application based on a limited dataset.

AR: We will further hint at the limitations emerged.

Pag.20, Lines 5-12:Authors recognize the major drawback of the present study. They would evaluate a new model-based approach to flood forecasting which does not re-quire the calibration task for the hydrological module, but the available dataset is not appropriate (due to the very limited size) to prove the added value of the proposed approach.

AR: Again, we are not seeking for added value, but for a useful tool that do not require calibration. The results with a limited set of data show us that we are on the right way.

References:

Addor, N., Jaun, S., Fundel, F., and Zappa, M.: An operational hydrological ensemble prediction system for the city of Zurich (Switzerland): Skill, case studies and scenarios, Hydrology and Earth System Sciences, 15, 2327–2347, doi:10.5194/hess-15-2327-2011, 2011.

Antonetti, M., Horat, C., Sideris, I. V., and Zappa, M.: Ensemble flood forecasting considering dominant runoff processes: I. Setup and application to nested basins (Emme, Switzerland), Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2018-118, in review, 2018.

Cane, D., Ghigo, S., Rabuffetti, D., and Milelli, M.: Real-time flood forecasting coupling different postprocessing techniques of precipitation forecast ensembles with a distributed hydrological model. The case study of may 2008 flood in western Piemonte, Italy, Nat. Hazards Earth Syst. Sci., 13, 211-220, https://doi.org/10.5194/nhess-13-211-2013, 2013.

Devoli, G., Tiranti, D., Cremonini, R., Sund, M., and Boje, S.: Comparison of landslide forecasting services in Piedmont (Italy) and Norway, illustrated by events in late spring 2013, Nat. Hazards Earth Syst. Sci., 18, 1351-1372, https://doi.org/10.5194/nhess-18-1351-2018, 2018.

Kobayashi, K., Otsuka, S., Apip, and Saito, K.: Ensemble flood simulation for a small dam catchment in Japan using 10 and 2 km resolution nonhydrostatic model rainfalls, Nat. Hazards Earth Syst. Sci., 16, 1821-1839, https://doi.org/10.5194/nhess-16-1821-2016, 2016.

Li, Z., Li, Y., Bonsal, B., Manson, A. H., and Scaff, L.: Combined impacts of ENSO and MJO on the 2015 growing season drought on the Canadian Prairies, Hydrol. Earth Syst. Sci., 22, 5057-5067, https://doi.org/10.5194/hess-22-5057-2018, 2018.

Picciotti, E., Marzano, F. S., Anagnostou, E. N., Kalogiros, J., Fessas, Y., Volpi, A., Cazac, V., Pace, R., Cinque, G., Bernardini, L., De Sanctis, K., Di Fabio, S., Montopoli, M., Anagnostou, M. N., Telleschi, A., Dimitriou, E., and Stella, J.: Coupling X-band dual-polarized mini-radars and hydro-meteorological forecast models: the HYDRORAD project, Nat. Hazards Earth Syst. Sci., 13, 1229-1241, https://doi.org/10.5194/nhess-13-1229-2013, 2013.

Philipp, A., Kerl, F., Büttner, U., Metzkes, C., Singer, T., Wagner, M., and Schütze, N.: Small-scale (flash) flood early warning in the light of operational requirements: opportunities and limits with regard to user demands, driving data, and hydrologic modeling techniques, Proc. IAHS, 373, 201-208, https://doi.org/10.5194/piahs-373-201-2016, 2016.

Zappa, M., Jaun, S., Germann, U., Walser, A., and Fundel, F.: Superposition of three sources of uncertainties in operational flood forecasting chains, Atmospheric Research, 100, 246–262, doi:doi:10.1016/j.atmosres.2010.12.005, 2011.