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

Authors replies to RC1 (Eric Gaume):

We want to thank Eric Gaume for his 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. This reply is in some parts identical with the already uploaded reply to reviewers 2, as some issues have been raised by both reviewers.

GENERAL ASSESSMENT

RC: The connection with real-world operational methods and application is a very positive aspect of the presented work. To my knowledge, particularly advanced methods are implemented operationally in Switzerland if compared to the rest of the world. Nevertheless, the proposed comparison is conducted for one single watershed (and one of its sub-watersheds but the results are not presented in the manuscript) and for a period of time of 3 months only (May to August 2016). This is by far insufficient to draw real convincing conclusions…… Likewise, a three months period seems far too short for a faithful evaluation of forecasting models and does not correspond to the general standards of the scientific publications. The results may be too dependent on some few if not a single flood event with no possibility of generalization. The authors themselves acknowledge in the discussion part of their manuscript that sparse data may be problematic (P 18 L6)…. The objectives and methods presented in the paper are correct, but the manuscript can hardly be published to my opinion unless a much larger data set in time and space (larger period of time and larger number of watershed is considered). The team seems to have access to reach data sets in Switzerland ; It is time consuming of course, but I do not see any reason why they could not conduct a large and necessary test and validation study based on the approach presented in the manuscript.

AR: This issue has been raised also by the reviewer 2 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 also 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. Even Switzerland cannot afford to generate re-forecasts of ensemble weather predictions, and only a limited time was available to compare the old atmospheric EPS system (COSMO-LEPS) with the new one (COSMO-E). As COSMOE is the future, we decided to focus on the available COSMOE data.

In the revised manuscript 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 both 2017 and 2018, this would not really beneficial since no severe flood event occurred (see also our detailed reply to reviewer 2). Summarizing, for the time being we are not able to generate "useful" additional runs of our event-based model RGM-PRO.

Here we are not looking for added value with respect to the traditional forecasts, 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.

For this study, we selected two different chains are, 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 extreme 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.

RC: The implementation of a rainfall-runoff model without calibration can only be evaluated if conducted on a significant number of watersheds – typically some tens.

AC: RGM-PRO has been introduced by Antonetti et al. (2017) with an analysis of 5 basins and 8 events. In Antonetti et al. (2018) three additional basins have been implemented and discussed. In this manuscript RGM-PRO is configured (not calibrated) for the Verzasca to

increase (by two, with the nested basin) the number of applications. The Verzasca basin is one of four basins where we could have performed such a benchmark with "traditional" operational forecasts, the other three being sub-areas of the Sihl river (Addor et al., 2011), for which have by far less published work on flash-flood than in case of the Verzasca.

This table has been also provided to the reviewers of our companion paper to show how these papers relate to our previous studies. The table will be included in the companion paper by Antonetti et a.l (NHESSD).

Paper	*Zappa et al.*	*Addor et al.*	*Liechti et al.*	*Antonetti et al.*	*Antonetti et al.*	*Antonetti et al.*	*Horat et al.*
Year	*2011*	*2011*	*2013*	*2017*	*2018*	*2018*	*2018*
Journal	*At. Research*	*HESS*	*HESS*	*Hydrol. Proc.*	*HESS*	*NHESSD*	*NHESSD*
Target areas							
Verzasca	X		X				X
Sihl		X					
Emme					X	X	
Other			X	X			
Topics							
Forecasting	X	X	X			X	X
Model development				X			
Uncertainty propagation	X		X		X	(X)	(X)
Intercomparison		X	X	(X)	(X)	X	X
Model/module							
PREVAH-HRU	X	X	X				X
RGM-PRO				X	X	X	X
RGM-TRD					X	X	
Rainfall forcing							
Intrepolated gauges	X	X	X		X		X
Combiprecip				X	X	X	X
COSMO-1						X	X
COSMO-2	X	X	X				(X)
COSMO-LEPS	X	X					(X)
COSMO-E						X	X
Weather radar nowcasting	X		X				
Frequency	continuous	continuous	events	events	events	events	events
Period	2007-2010	2007-2009	2007-2010	2005-2016	2005-2016	2016	2016
Analyses							
NSE/KGE	NSE			KGE	KGE	NSE/KGE	
Brier/ROC/FAR/RankHist	(X)	X	X			X	X
MonteCarlo	X		(X)	X	X	X	(X)
Other	SWAE				ANOVA		

Zappa et al. (2011) is our benchmark paper on uncertainty propagation

Addor et al.. (2011) is our reference work on verification of deterministic and ensemble forecasts (with the Breir score as main metric to discriminate between deterministic and probabilistic forecasts).

Liechti et al. (2013) focuses on flash-flood nowcasting with advanced weather radar products

Antonetti et al. (2017) introduces RGM-PRO

Antonetti et al. (2018, HESS) evaluate structures and configurations of RGM-PRO in the Emme catchment

Antonetti et al. (2018, NHESSD) first apply RGM-PRO in forecasting mode for the Emme catchment and is our first study with COSMO-E/COSMO-1

Horat et al. (2018, NHESSD) applies RGM-PRO in forecasting mode for the Verzasca catchment and compare its quality with our current operational model as forced by COSMO-E/COSMO-1.

RC: 1) The authors mostly refer to their own works. Indeed, interesting and innovative methods are implemented in Switzerland to forecast flash floods. But it would important also to cite works conducted in other countries and by other teams on the same issue at least in the introduction of the manuscript to show the originality of the proposed approach. Flash flood forecasting has been an active field of research in the recent years.

AR: The reviewer is right. The introduction of this "Part II" was the most difficult section to write, as we already review recent work on flash flood in "Part I". We did not want to replicate great parts of the introduction from "Part I" here and thus focussed on introducing the state-of-the-art of our applications in the Verzasca river. We will try to re-formulate the introduction and include a better-balance between own work, previous work and the review presented for "Part I".

RC: 2) The manuscript refers in many places to a companion paper and to supplementary materials. This is frustrating for the readers since some important information is not provided in the manuscript such as the implementation of the "process based" model (what are the input variables, how are the values of the parameters of the model fixed) or the results obtained for the Pincascia sub-watershed. Supplementary material is interesting but a manuscript must be to a certain extent self-sufficient and contain at least the basic information needed for the interpretations and the results that are commented and interpreted.

AR: We are of the opinion that "Part I" and "Part II" are together "self-sufficient" but agree that for a reviewer providing the review of "Part II" only, some frustration might arise, because some of the methods have been detailed in "Part I" and previous publications. The document the reviewer assessed is not intended to introduce RGM-PRO (see Table on the previous page). We compare here two published models, one of them do not require calibration and uses the parameterization presented in Antonetti et al. (2017). The PREVAH-HRU chain is (beside the COSMOE/COSMO1 forcing) identical with the chain first presented in Zappa et al., (2011). In the revised version we will try to make this manuscript more self-sufficient without replicating what shown in other papers. As far as the use of supplementary material is concerned, we selected to follow the journals general wish of manuscripts with a reduced number of figures and moved figures showing the same for another area to the supplementary material, while keeping both basins in the synthesis presented in Figure 8. In the revised manuscript we will include the results of Figure 6 and 7 for the Pincascia basin in the main manuscript.

RC: 3) Brier scores are used to compare deterministic and probabilistic forecasts. I know that some other papers did the same, but this comparison is not appropriate. Indeed, a Brier score can be computed in both cases, but do not measure exactly the same things and can therefore not be directly compared. Forecasts must be combined with a utility function and evaluated in a decision making context for a proper and rigorous comparison.

AR: The Brier Score (BS) is our link to previous studies where we are well confident that BS is a very efficient way to discriminate between the skill of deterministic and probabilistic forecasts. We will elaborate on this statement in the revised manuscript.

COMMENTED DOCUMENT

RC: Usually, no refrences are cited in the abstract of a paper

AR: We will remove the citation to the companion paper from the abstract.

RC: This seems to be a too short period for a real evaluation of forecasting model even if a 2-year flood has been observed during this Summer.

AR: See replies to the general comments

RC: Such a conclusion cannot be drawn based on two application examples only. Wider applications of the proposed methods would be needed to confirm or invalidate this impression.

AR: See replies to the general comments

RC: Again, I have here some doubts and in any case no general conclusion ca be drawn from the presented limited test case. Several previous studies, including the so-called DMIP experiment (Distributed model intercomparison program) did not lead to the conclusion that "processed based" models do have better performances than calibrated conceptual models. On the contrary...

AR: We are not aiming at declaring that RGM-PRO without calibration is better that the (process-based, calibrated) PREVAH-HRU model with very conceptual runoff generation module. We are happy to see that they show similar quality in a period including a flood with 2-years return period. Thanks for mentioning DMIP. We will include this in the introduction and discussion.

RC: Why such a detailled emphasis on Doswell's work ?

AR: We will condense this part of the introduction and add work of other authors.

RC: Forecasted : they are some spelling errors in the manuscript that need to be corrected.

AR: We corrected this occurrence and will look for other typos as suggested by the reviewer.

RC: The literature review is mostly citing works conducted in Switzerland by the authors' team. It could be enlarged significantly.

AR: See replies to the general comments

RC: I am not found of the term process or physically based since the implementation scale of the models is generally much too coarse to enable a detailed description of the hydrological processes. More-over, even the process based model generally need some calibration.

AR: RGM_PRO is a dominant runoff process based module for the process of runoff generation. Parameters can be set a priori basing on calibration of sprinkling experiments under different conditions as introduced in detail by Antonetti et al. (2017) and as illustrated in "Part I" of this manuscript. We will expand this section in order to give a better orientation to the readers that are not familiar with our previous study.

RC: Is an hourly time step really suited for such a small watershed ?

CombiPrecip would be available in 10 minutes step. The disaggregation of COSMO-forecasts into 10 minutes field goes beyond the scope of this paper. According to our experience and for basins in the order of size of the Verzasca are one hour forcing data sufficient. The model internal integration time step is of 10 minutes.

RC: Please develop the acronym (FOEN)

AR: Already defined on Page 3.

RC: Develop the acronym (SPPT)

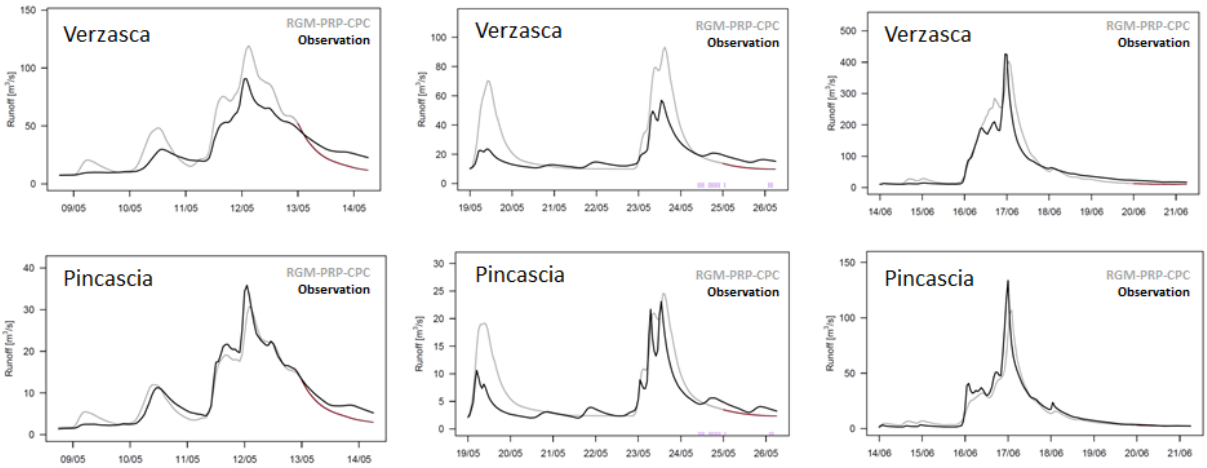
AR: Section 3.2.2 will be expanded as suggested by reviewer 2. All acronyms will be defined.

RC: 1 alert every 5 days ! The number of detections and False alarms could be indicated here.

AR: Will be done.

RC: I am not found of the reference to supplementary material. An article should be self-sufficient. If the information is important for the analysis, it should be provided in the paper.

AR: We thank the reviewer for this amendment. A figure will be added and text will be included to close this gap of information.



New Figure: Flood hydrographs of reference simulation by RGM-PRO as forced by CombiPrecip (cpc) for three events in the analysis period. Top panels: Verzasca basin. Bottom panels: Pincascia river.

RC: The references to the companion paper are too frequent. Again, the manuscript should be self-sufficient and contain the necessary information and results.

AR: This sentence can actually be removed and will be removed.

RC: The definition of BS could be recalled. I am wondering if BS values computed for deterministic models with binary results (1 exceedance and 0 non-exceedance) and for probabilistic models (probability of exceedance) are comparable ? Binary models may provide systematically larger BS values...

AR: … and thus less skill for users needing taking action basing on a threshold. The question is if an ensemble model with coarser resolution is more useful than a high resolution deterministic mode. We will surely expand on the BS issue to clarify our reason for using it.

RC: the signification of symbol o should be explained. Add a legend to this equation.

AR: Will be done.

RC: 15:12:16 BS and BSS can be computed for both types of forecasts. I wonder if the obtained values can directly be compared. Probably not).

RC: Again : include important results in the manuscript.

AR: This one will stay in the supplementary material.

RC: The bootstrappping procedure must be better explained. A few sentences can be added. The reader cannot understand what exactly is resampled.

AR: Events are removed from the available sample and the score are recalculated. This allows for quantification of uncertainty stemming from the choice of events. Clarification will be included in the revised manuscript.

RC: It is important to mention somewhere in the manuscript how many events are included in each analysed sample.

AR: The number of cases per each rainfall intensity evaluated will be provided.

RC: Sign of larger discharge values produced by the process-based model. The balance between POD and FAR must be better justified. What are the operational needs. How would end-users judge both results?

AR: The traditional model is quite lumped with respect to runoff generation description and thus tends to smooth intense local precipitation in terms of output simulated discharge. The process based RGM-PRO accounts for local impervious areas with rapid transformation of rainfall into discharge and thus yields generally more "flashy" hydrographs. The fact of having high POD is surely an advantage, but high FAR is also not nice from an user perspective. In the end each user should decide his risk profile under consideration of costs of action and remaining risk that he can accept. For this case we have no user to ask, but in case of the Sihl forecast presented in Addor et al. (2011) the user is surely interested of having no missed events and is ready to take action even if an event would not occur in the end. We will elaborate these thoughts in the manuscript.

RC: Some explanation of the contrasted behaviour of both models could be interesting for the reader.

AR: We will elaborate this remark.

RC: Or even much more, since results for larger lead times are not shown. The stability of the skill score with lead times is a surprising result a deserves some comments.

AR: COSMO1 lead time do not go beyond the shown range. COSMOE does and we show the results for the lead times beyond hour 29 in Figure 8.

RC: This is what is generally expected

AR: Yes, whereby in some cases ensembles need some hours to develop and best skill is obtained between 24 and 48 hours lead time.

RC: That is less systematic

AR: Yes.

RC: This very surprising result should not only be described but analysed...

AR: We think that this is due to the fact, that Addor et al. (2011) do not evaluate at sub-daily scale and thus the "constant" skill in the first 36 hours is not intelligible. Liechti et al. (2013) do not use ensemble EPS, and thus the increase of skill in the first days in not shown for an ensemble system. Such behaviour has been described for the Verzasca in Zappa et al. (2013, Figure 3.), where maximum skill (ROCa) was between day 3 and 4.

RC: Yes, the presented results may be extremely dependent on the specificities of the analyzed sample that is of limited size. Is it really possible to draw valuable general conclusions ?

AR: Yes, it is, because as already stated both chain "suffers" from a limited sample size.

RC: Again a strange results. How many events control the criteria values for large thresholds ?

AR: Very few.

RC: This may depend on the antecedent moisture conditions

AR: The reviewer might be right, but as already discussed, we think that the lumped structure in runoff generation description of PREVAH-HRU leads to a general "inertia" during the wetting phase.

RC: Unclear sentence

AR: Re-formulation: "Furthermore, the process-based forecasting chains react with less delay to rainfall input, leading to higher peaks in runoff but also larger uncertainties when applied for ensemble forecasting. Although the use of information about DRP decreases the hydrological model parameter uncertainty, as found by Antonetti et al. (2016b), it does not decrease the total uncertainty when ensemble precipitation forecasts are used. In other words the fully distributed DRP-approach amplifies the spread originating from different members of a precipitation ensemble, while the semi-distributed PREVAH-HRU approach strongly smooths such differences.

RC: The authors are comparing their work to their previous works... Is this interesting for a broader audience ?

AR: If the reader is interested on performance of an uncalibrated model as compared to the one of a calibrated model then yes. We will re-arrange sections 6.3 to 6.6 as suggested by reviewer 2.

RC: Please provide some explanation to this observed result

AR: "Sample size", but maybe also skill of the DRP approach in case of larger flood peaks.

RC: Readers that are not familiar with the author's work can hardly follow. What is the aim of comparing the two published results if according to the authors the methods are too different to really enable a comparison ?

AR: Our original formulation might suggest that we are looking for the "best" approach, while as stated before, we want to show that comparable results can be achieved without calibration. We will adapt this sentence in order to be compliant with our goals.

RC: Again, I am not sure that both BSS values can be compared. Discrepancies between forecasts and observations lead to BS contributions equal to 1 for discrete models, while contributions are always much lower than one (difference at a power two) for ensemble approaches.

AR: We see the point of the reviewer and can suggest to this issue this blog on HEPEX: https://hepex.irstea.fr/how-can-the-brier-score-know-my-inner-thoughts/ Even if we will not make a "Brier" paper out of it we will add more thoughts on its use for comparing deterministic and ensemble forecasts.

RC:, that is a major issue, and the data set used in this study is extremely sparse (3 months). Are the obtained results really meaningful and worth an interpretation. I really have major doubts.

AR: This re-iterates previous comments on this issue.

RC: Yes, that is also very true and would provide additional elements for the interpretation of the results. The authors have the data, why did they not conduct this analysis and include it in the manuscript ?

AR: As also replied to reviewer 2, the output diagnostic from COSMOE and COSMOE is automated and re-configuration would have been beyond the scope of the paper. In the meantime a paper on the verification of COSMOE has been presented by Klasa et al. (2018)

RC: Is this a really significant result or does it illustrate the consequence of the dependence of the result to the limited sample (sampling variability) ?

AR: Significant at a very preliminary level.

AR: This is a pure speculation at this stage. This hypothesis needs to be confirmed on a larger set of test cases before it can really be formulated.

AR: We will formulate the sentence to render its speculative character.

AR: No model can be really described as process-based or physically-based. Ideally the term should be in quotation marks...

AR: Will be done.

RC: Perhaps "with no calibration" or "implemented without being calibrated", but to my experience and the experience of most hydrological modelers, every model benefits from calibration...

AR: We agree that each model can benefit from calibration, when observations are available. RGM-Pro process-based reaction to precipitation has been calibrated against sprinkling experiments (Antonetti et al., 2017). This relation is now used as "universal" parameterization for any application of RGM-PRO. The mapping of the areas where

processes occur determine how specific catchments reacts. We mapped the Verzasca and applied the RGM-PRO parametrization without any further calibration.

RC: intensely

AR: Thanks for the correction.

RC: Yes, this is my major critic. I read this sentence as a clear understatement. Is not fully appropriate or fully inappropriate in the sense that the data set is far insufficient to draw clear conclusions.

AR: See our previous replies.

RC: Some elements are missing explaining the implementation principles of the new approach. The new model has also parameters. What source of data has been used to fix the value of these parameters ?

AR: See previous replies (sprinkling experiments) and Antonetti et al. (2017).

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.

Klasa C, Arpagaus M, Walser A, Wernli H. An evaluation of the convection-permitting ensemble COSMO-E for three contrasting precipitation events in Switzerland. Q J R Meteorol Soc. 2018;144:744–764. https://doi.org/10.1002/qj.3245

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.

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.