Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-102, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 1 March 2017

Hydrology and Earth System Sciences
Discussions

Comment on "Can assimilation of crowdsourced data in hydrological modelling improve flood prediction?" by Mazzoleni et al. (2017)

Daniele P. Viero[1]

[1]Department of Civil, Environmental, and Architectural Engineering, University of Padova, via Loredan 20, 35131, Padova (Italy).

Correspondence to: Daniele P. Viero (daniele.viero@unipd.it)

Abstract. In their recent contribution, Mazzoleni et al. (2017) investigated the integration of crowdsourced data (CSD) in hydrological models to improve the accuracy of real-time flood forecast. They showed that assimilation of CSD improves the overall model performance in all the considered case studies. The impact of irregular frequency of available crowdsourced data, and that of data uncertainty, were also deeply assessed. However, it has to be remarked that, in their work, the Authors used synthetic (i.e., not actually measured) crowdsourced data, because actual crowdsourced data were not available at the moment of the study. This point, briefly mentioned by the authors, deserves further discussion. In most real-world applications, rainfall-runoff models are calibrated using data from traditional sensors. Typically, CSD are collected at different locations, where semi-distributed models are not calibrated. In a context of equifinality and of poor identifiability of model parameters, the model internal states can hardly mimic the actual system states away from calibration points, thus reducing the chances of success in assimilating real (i.e., not synthetic) CSD. Additional criteria are given that are useful for the a-priori evaluation of crowdsourced data for real-time flood forecasting and, hopefully, to plan apt design strategies for both model calibration and collection of crowdsourced data.

1 Introduction

The availability of hydrometric data, collected by active citizens in the course of severe flood events, offers a new, unexpected chance to improve real-time flood forecasts. In pioneering applications, crowdsourced data (CSD) collected in the upper part of a basin were assimilated into adaptive hydrologic models to reduce the uncertainty in forecasting flood hydrographs at downstream sections (Mazzoleni et al., 2015). In a recent work, Mazzoleni et al. (2017) paid particular attention to the issues of data uncertainty and irregular arrival frequency of CSD. Their results showed that assimilation of CSD improves the overall model performance in all the case studies they considered. They also showed that the accuracy of CSD is, in general, more important than their arrival frequency.

However, there is a crucial aspect that has to be remarked. In their work, the Authors used synthetic (i.e., not actually measured) CSD, because real streamflow CSD were not available at the moment of the study. The Authors warned about this aspect by stating that "*the developed methodology is not tested with data coming from actual social sensors. Therefore, the*

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-102, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 1 March 2017

Hydrology and Earth System Sciences
Discussions

conclusions need to be confirmed using real crowdsourced observations of water level". This point deserves further discussion, as the use of synthetic data led them to disregard a subtle, yet significant, limitation inherent in the use of CSD in real-time flood forecasting. The problem involves equifinality (i.e., uncertainty in model parameters and internal states, Beven, 2006) that characterizes hydrologic, semi-distributed (and over-parametrized) models.

After the critical work by Beven (1989), detailed investigations were carried out about the complexity a model needs to simulate rainfall-runoff process. Several studies indicated that the information content in a rainfall-runoff record is sufficient to support models of only very limited complexity (Jakeman and Hornberger, 1993; Refsgaard, 1997). This implies that distributed, or semi-distributed, hydrologic models are seldom calibrated. Rather, they are commonly over-parametrized. As a typical example, a semi-distributed rainfall-runoff model may provide accurate predictions of the outflow discharge at the closing section and, at the same time, it can fail to correctly model the relative contribution of upstream tributaries. To limit problems related to over-parametrization, also the internal states of a distributed model have to be calibrated (Sebben et al., 2012; Viero et al., 2014), and not only the outflow at the closing section.

Strictly speaking, and bearing in mind that one can get the correct answer for the wrong reason (Loague et al., 2010), a semi-distributed model can be said calibrated only at the calibration points. This caveat has important consequences also on data assimilation and models updating.

In general, data assimilation techniques are used to update model input, states, parameters, or outputs based on new, available observations (Refsgaard, 1997). Assimilation of CSD may improve the performance of a forecasting model inasmuch as assimilated data contribute in updating (i.e., in correcting) the internal states of the model. It must be observed that crowdsourced data typically refers to internal states of the model, since input and output data commonly corresponds to location provided with traditional physical sensors. For updating to be successful, available data must be substantial and accurate (as well debated by Mazzoleni et al., 2017), but further requirements must be met. Indeed, data assimilation is successful if the model can correctly predict, at the same time, both the main output and the internal states of the system. At least, the model have to describe well the real system states (i.e., must be properly calibrated) at every location in which crowdsourced data are collected. Accordingly, crowdsourced data must be collected in correspondence of the control points of the models (i.e., those used to calibrate the model).

Therefore, beside the key points identified by Mazzoleni et al. (2017), not only data, but also the model has to match specific requirements for data assimilation to be successful. This issue is certainly relevant for the case study of the Bacchiglione River, for the reason reported in the following.

2 Specific comments

In this Section, the focus is on the fourth case study presented in Mazzoleni et al. (2017), in which synthetic (i.e., not actually recorded) crowdsourced data (CSD) were used to improve the performance of a semi-distributed hydrological model of the Bacchiglione catchment closed at Ponte degli Angeli, Vicenza (Italy).

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-102, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 1 March 2017

Hydrology and Earth System Sciences
Discussions

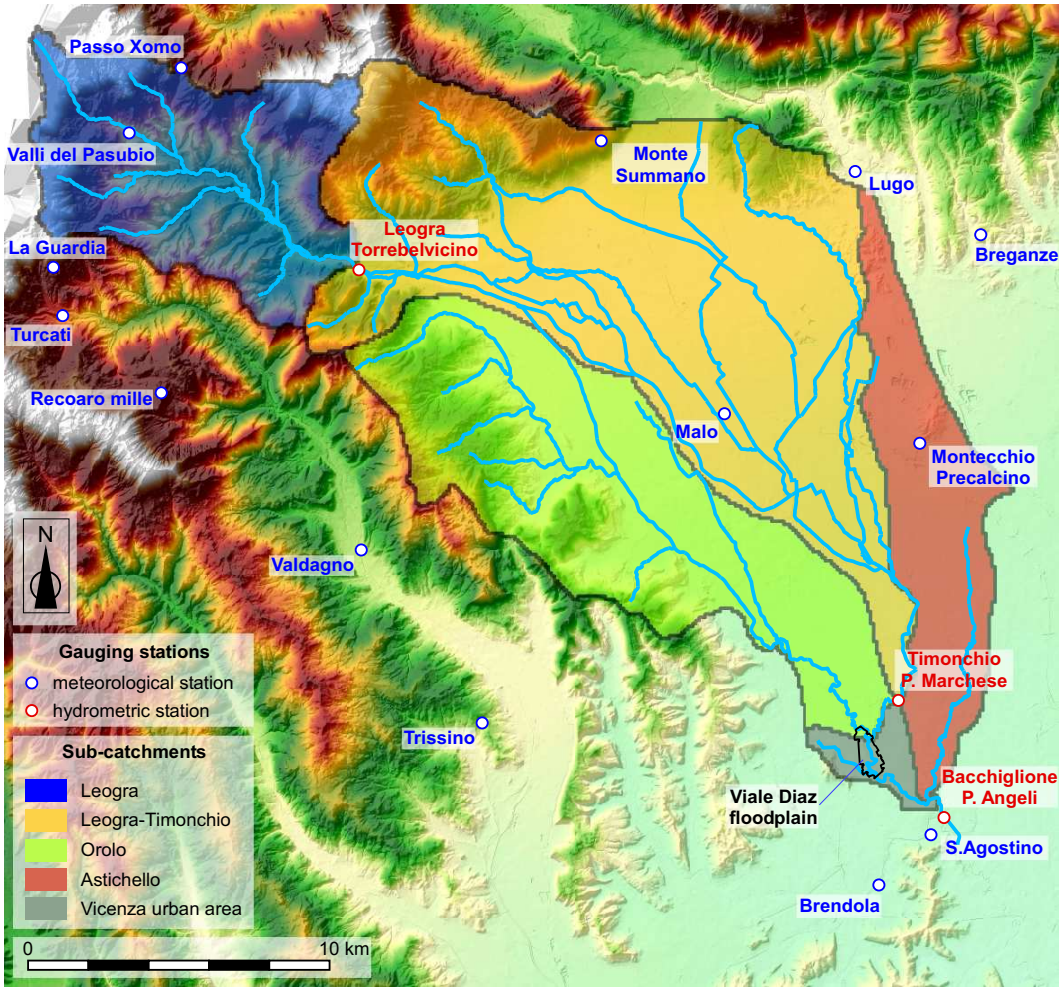

Figure 1. The catchment of the Bacchiglione River closed at Ponte degli Angeli, Vicenza (Italy).

2.1 The Bacchiglione catchment closed at Ponte degli Angeli (Vicenza)

The catchment of the upper Bacchiglione River, closed at Ponte degli Angeli in the historical centre of Vicenza (Fig. 1), is located in the north of the Veneto Region, a plain that is fringed by the Alpine barrier at a distance of less than 100 km to the north of the Adriatic Sea (Barbi et al., 2012).

With regard to the precipitation climatology, the southern part of this plain is the drier, with approximately 700–1000 mm of mean annual rainfall, whereas more than 2000 mm are measured close to the pre-alpine chain. Obviously, these differences are mainly related to the mountain barrier and its interaction with southerly warm and humid currents coming from the Mediterranean Sea (Smith, 1979). Indeed, the topography of the region rises from the southern plain at about 30 m above sea level (a.s.l.) to about 1500–2200 m a.s.l. in the first orographic barrier, the pre-alpine chain, and then further to the north to the

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-102, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 1 March 2017

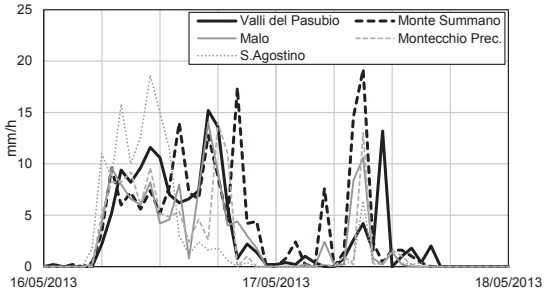

Figure 2. Hourly rainfall rates for the storm event of May 2013, 16-18.

Dolomites, a mountain massive that peaks at over 3000 m a.s.l. In the northern part of the Bacchiglione catchment, the terrain elevations raise from 250 to 1'000 m a.s.l. in less than 1 km, with slopes up to 70%.

A significant portion of the annual rainfall often concentrates into very short periods of time in the form of what often turns out to be an extreme event with deep convection playing a central role (Barbi et al., 2012; Rysman et al., 2016). As a consequence, severe flooding event have threatened agricultural and urban areas in the recent years (e.g. Viero et al., 2013; Scorzini and Frank, 2015).

A comparison of hourly rainfall rates measured at the four meteorological gauging stations of Valli del Pasubio, Monte Summano, Malo, Montecchio Precalcino, and S.Agostino (Fig. 1) is reported in Fig. 2 for the storm event of 16-18 May 2013 (data provided by the Regional Agency for Flood Protection of the Veneto Region, ARPAV). The spatial and temporal variability of the rainfall fields is apparent.

Many meteorological model are unable to provide accurate and reliable quantitative precipitation estimates (QPE) for the upper Bcchiglione catchment, due to both insufficient spatial and temporal resolution, and to the actual complexity of this environment. An example ot this inadequacy is given, for instance, by Fig. 13 in Mazzoleni et al. (2017). The discharge simulated using forecasted input is very different from that obtained using recorded rainfall, showing significant time shift and errors between 25 and 50% at the flood peak (and up to 90% if considering synchronous data).

From an hydraulic point of view, the upper Veneto plain is a highly populated and urbanized area, with extremely complex drainage and irrigation networks. Within this plain, the Bacchiglione River and all its tributaries are provided with relatively high levees (Viero et al., 2013), which prevent the exchange of water from inside to outside the riverbed (and vice-versa) when the inner water levels are relatively high. As a consequence, the minor channel networks are not always allowed to deliver their drainage water towards the nearest tributary, i.e., the inflow points along the main river reaches change during a flood event depending on the instantaneous water level within the river. This occurrence change the network connectedness which, in turn, leads to different mechanisms of hydrologic response in the overall catchment.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-102, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 1 March 2017

Hydrology and Earth System Sciences
Discussions

Just upstream of the City of Vicenza, a floodplain of about 1 km^2 is flooded when the flow rate in the Bacchiglione exceeds ~ 160 m^3/s. Since about $2 \cdot 10^6$ m^3 of water can be temporarily stored in this area, a significant flood attenuation can be produced, particularly in case of floods with a steep rising limb (which is often the case).

Clearly, such a system is highly non-linear. Nonetheless, significant parts of the Bacchiglione catchments are poorly monitored, and the remaining parts are completely unmonitored. The Leogra subcatchment (blue shaded area in Fig. 1) is provided with a pressure-transducer for the measure of water level at Torrebelvicino (Fig. 1). A rating curve, derived from theoretical considerations, is available for this cross-section. Its reliability is clearly low, since no instrumental measures of flow discharge are available for this site. The Leogra-Timonchhio subcatchment (orange shaded area in Fig. 1) is monitored by a ultrasonic stage sensor operated by ARPAV; Located in Ponte Marchese, just upstream of the confluence with the Orolo River, it is not provided with any rating curve. Available flow rate measures at Ponte Marchese refers only to low hydraulic regimes, and show great variability due to the operations of a hydroelectric power plant located just downstream of Ponte Marchese. The Orolo River (green shaded area in Fig. 1), with a discharge capacity of more than one third of the Bacchiglione at Ponte degli Angeli, is one of its major tributaries. The catchment of the the Orolo River leans against a ridge, which increases the spatial variability of precipitation fields. Unfortunately, not only this area is completely uncovered by meteorological gauging stations, but also no hydrometric gauging station are present along the reach of the Orolo River. Similarly to the Orolo, the Astichello catchment (red shaded area in Fig. 1) is unmonitored and, due to backwater effects, significant areas adjacent to the Astichello are flooded when water levels in the Bacchiglione are relatively high. Hence, the discharge that effectively flows from the Astichello into the Bacchiglione River may significantly reduced depending on the water stage within the main course of the Bacchiglione River.

Attention must be paid to the fact that the three major tributaries (Orolo, Timonchio, and Astichello) meet just upstream of the closing section of Ponte degli Angeli (Fig. 1), making it difficult to estimate the actual contribution of each single tributary to the total streamflow correctly. By looking at the tree-like structure of the drainage network (Rodríguez-Iturbe and Rinaldo, 2001) in an electrical analogy, the major tributaries of the Bacchiglione are in fact “conductors in parallel”.

Finally, the lower part of the Bacchiglione basin, North of Vicenza, includes a vast groundwater resurgence zone, in which it’s difficult to assess both the actual contribution of resurgence to the Bacchiglione streamflow (up to ~ 30 m^3/s) and the time-variable behaviour of soil moisture.

Certainly, given the irregular topography of the catchments, the heterogeneity of the landscape, and the complexity of the hydraulic network, it can be stated that the catchment of Bacchiglione is poorly monitored.

2.2 The semi-distributed model of the Bacchiglione catchment

In catchments like that of Bacchiglione, for all the reasons reported in the previous section, the accurate prediction of flood hydrographs by performing continuous time simulations is unquestionably a hard task (Anquetin et al., 2010).

Sensibly, the semi-distributed model used in Mazzoleni et al. (2017) was calibrated by minimizing the root mean square error between observed and simulated values of water discharge only at the Ponte degli Angeli, which is the only hydrometric station provided with a reliable rating curve. The semi-distributed model, although explicitly representing the hydrological processes

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-102, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 1 March 2017

within the main subcatchments, has to be intended as a lumped model from a practical standpoint, since the discharge in Ponte degli Angeli is its only control point.

Therefore, no matter the accuracy of the model in forecasting flood hydrographs in Ponte degli Angeli, little can be said about the accuracy of the same model in describing the internal states of the system, such as the streamflow along the upstream tributaries. This limitation has to be ascribed to uncertainty in precipitation fields, to the paucity of (reliable) flow rate data upstream of Vicenza, and to inherent limitations of the model itself.

Indeed, it has to be remarked that the semi-distributed hydrologic model used by Mazzoleni et al. (2017) accounts for flood propagation by means of a Muskingum–Cunge model that considers rectangular river cross-sections for the estimation of hydraulic radios, wave celerities, and other hydraulic variables (Todini, 2007). Accordingly, the effects exerted by the "Viale Diaz" floodplain, which acts as a sort of in-line natural flood control reservoir on flood propagation, can not be properly accounted for. This means that, if the flood hydrograph is correctly modelled at Ponte degli Angeli, it is not correctly modelled upstream of the Viale Diaz floodplain (and vice-versa).

2.3 The use of CSD in a context of equifinality

In the work by Mazzoleni et al. (2017), the synthetic hourly crowdsourced data (CSD) of streamflow are the result of the model itself. Indeed, synthetic CSD were calculated by forcing the hydrological model of the Bacchiglione catchment with measured precipitation recorded during the considered flood events (post-event simulation). As a matter of fact, these data are representative of the actual model internal states of the best-fit scenario.

Importantly, the synthetic CSD used by Mazzoleni et al. (2017) in the Bacchiglione case study do not refer to calibration points of the model. This aspect can be seen as a peculiarity of crowdsourced data, whose natural purpose is to enhance (rather than replace) data from traditional sensors. Indeed, historical data recorded by traditional sensors are first used to calibrate a model; then, in real-time mode, the same sensors provide data both to force the model and to update the model states (e.g. Ercolani and Castelli, 2017); moreover, the reliability of data from traditional sensors outperform that of CSD.

The Author claimed that the synthetic CSD they used are realistic. For the Bacchiglione case study, recalling the global picture given in Sections 2.1 and 2.2, and that the semi-distributed model was calibrated only at closing section of Ponte degli Angeli, this statement is at least questionable. Indeed, for synthetic streamflow CSD to be realistic, two specific requirements have to be met: *i)* a reliable rating curve must be available for the cross sections where hydrometric CSD are recorded, and *ii)* the model has to be calibrated at these locations. Unfortunately, none of these requirements are met for the Bacchiglione River. The first issue (i.e., lack of rating curves) was assessed inasmuch the Authors considered different degree of uncertainty in streamflow CSD. In this way, they accounted for, e.g., measuring errors and inaccuracy in rating curves. However, nothing was said (nor can be said) about the model performance at locations where CSD are collected, since these locations do not corresponds to calibration points. Here, the model predictions are likely biased but, contrarily to Mazzoleni et al. (2016), this aspect was not accounted for in Mazzoleni et al. (2017).

What can occur if, due to over-parametrization, the model badly reproduces the actual states at the CSD locations? In this case, the true crowdsourced data don't match the internal model states needed to produce an accurate prediction of the flood

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-102, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 1 March 2017

hydrograph at the downstream section. Their assimilation into the model can even lead to worse results than no assimilation at all or, at least, to fewer benefits than expected.

As warned by Dee (2005) and by Liu et al. (2012), great care should be taken in assimilating data if systematic biases or phase errors in the data or model exist, since the optimality of the data assimilation techniques is realized only if the observations and the models are not biased in the mean sense.

This observation is particularly important given that the results of the study by Mazzoleni et al. (2017) pointed out that the model performance is more sensitive to the accuracies of CSD than to the moments in time at which the streamflow CSD become available. Be careful that here, given the characteristics of CSD used by the Authors, "accuracy of CSD" implies a close similarity between the true crowdsourced data and the internal states of the model.

This problem is of general interest, and not limited to the study by Mazzoleni et al. (2017). Actually, the complexity of catchments, the relatively paucity of data, and the over-parametrization of semi-distributed rainfall-runoff models are likely the rule rather than the exception.

Therefore, the main aim of this comment is to warn about the subtle drawback hidden behind the (bad) practice of using traditional and crowdsourced data, recorded at different locations, disjointly; the former to calibrate (semi-)distributed models and to force them in real-time, the latter only to update the model states in operational forecasting. But the same problem, due to equifinality of (semi-)distributed models, could emerge due to a similar, incorrect use of only traditional data.

3 Summary

The approach proposed and investigated by Mazzoleni et al. (2017), based on the use of crowdsourced data (CSD) to improve real-time flood forecasts, is in general valuable, and shows a promising way to improve the accuracy of hydrological predictions using non-traditional information, which now active citizens and new technologies make available to hydrologists.

However, it has to be remarked that the correct description of the physical rainfall-runoff processes has to face actual limitations ascribed to the paucity of forcing data, to the complexity of real physical environments, and to the lacks in model structure and parametrization. As a consequence, rainfall-runoff models such as that used in Mazzoleni et al. (2017) can provide quite reliable predictions at locations where calibration is performed (i.e., control points), and still provide unacceptably wrong prediction of internal system states at the same time (e.g., discharge in ungauged tributaries).

In this context of equifinality (Beven, 2006), measured data that do not refer to calibration points of (semi-)distributed models are likely biased for data assimilation purpose (actually, at these locations, it is the model states that are biased rather than the measured data!). The performance of model updating can be substantially lower than expected when assimilating biased data (e.g., Dee, 2005; Liu et al., 2012). In other words, the assimilation of real (i.e., not synthetic) streamflow data referring to a poorly parametrized subcatchments or tributary can lead, in principle, to even worse model prediction than no assimilation at all.

The problem can arise due to the disjoint use of traditional and crowdsourced data that refer to different locations, with the former used to calibrate a (semi-)distributed model, and the latter used only in real-time model updating.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-102, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 1 March 2017

A pragmatic, operative recommendation is the collection of crowdsourced data for a suitable test period, to verify the model ability in describing the system states correctly at the locations in which CSD are collected, and possibly to update the model calibration using all the available data.

As a final remark, in order to take the maximum advantage in term of accurate and reliable real-time flood forecasts, both modellers and environmental agencies should account in a comprehensively manner for the characteristics of the physical system, for the model structure and parametrization, for the design of the sensor network, and for data to be used both in calibration and in operational mode.

References

Anquetin, S., Braud, I., Vannier, O., Viallet, P., Boudevillain, B., Creutin, J.-D., and Manus, C.: Sensitivity of the hydrological response to the variability of rainfall fields and soils for the Gard 2002 flash-flood event., J. Hydrol., 394, 134–147, doi:10.1016/j.jhydrol.2010.07.002, 2010.

Barbi, A., Monai, A., Racca, R., and Rossa, A.: Recurring features of extreme autumnal rainfall events on the Veneto coastal area, Nat. Hazard. Earth Sys. Sci., 12, 2463–2477, doi:10.5194/nhess-12-2463-2012, 2012.

Beven, K.: Changing ideas in hydrology: the case of physically based model., J. Hydrol., 105, 157–172, doi:10.1016/0022-1694(89)90101-7, 1989.

Beven, K.: A manifesto for the equifinality thesis., J. Hydrol., 320, 18–36, doi:10.1016/j.jhydrol.2005.07.007, 2006.

Dee, D.: Bias and data assimilation., Q. J. R. Meteorol. Soc., 131, 3323–3343, doi:10.1256/qj.05.137, 2005.

Ercolani, G. and Castelli, F.: Variational assimilation of streamflow data in distributed flood forecasting, Water Resour. Res., 53, doi:10.1002/2016WR019208, 2017.

Jakeman, A. and Hornberger, G.: How much complexity is warranted in a rainfall-runoff model?, Water Resour. Res., 29, 2637–2649, doi:10.1029/93WR00877, 1993.

Liu, Y., Weerts, A., Clark, M., Hendricks Franssen, H.-J., Kumar, S., Moradkhani, H., Seo, D.-J., Schwanenberg, D., Smith, P., van Dijk, A., van Velzen, N., He, M., Lee, H., Noh, S., Rakovec, O., and Restrepo, P.: Advancing data assimilation in operational hydrologic forecasting: progresses, challenges, and emerging opportunities, Hydrol. Earth Syst. Sci., 16, 3863–3887, doi:10.5194/hess-16-3863-2012, 2012.

Loague, K., Heppner, C., Ebel, B., and VanderKwaak, J.: The quixotic search for a comprehensive understanding of hydrologic response at the surface: Horton, Dunne, Dunton, and the role of concept-development simulation., Hydrol. Process., 24, 2499–2505, doi:10.1002/hyp.7834, 2010.

Mazzoleni, M., Alfonso, L., Chacon-Hurtado, J., and Solomatine, D.: Assimilating uncertain, dynamic and intermittent streamflow observations in hydrological models., Adv. Water Resour., 83, 323–339, doi:10.1016/j.advwatres.2015.07.004, 2015.

Mazzoleni, M., Alfonso, L., and Solomatine, D.: Influence of spatial distribution of sensors and observation accuracy on the assimilation of distributed streamflow data in hydrological modelling., Hydrolog. Sci. J., doi:10.1080/02626667.2016.1247211, 2016.

Mazzoleni, M., Verlaan, M., Alfonso, L., Monego, M., Norbiato, D., Ferri, M., and Solomatine, D.: Can assimilation of crowdsourced data in hydrological modelling improve flood prediction?, Hydrol. Earth Syst. Sc., 21, 839–861, doi:10.5194/hess-21-839-2017, 2017.

Refsgaard, J.: Parametrisation, calibration and validation of distributed hydrological models., J. Hydrol., 198, 69–97, doi:10.1016/S0022-1694(96)03329-X, 1997.

Rodríguez-Iturbe, I. and Rinaldo, A.: Fractal river basins: Chance and self-organization, Cambridge University Press, Cambridge, UK, 2001.

Rysman, J.-F., Lemaître, Y., and Moreau, E.: Spatial and temporal variability of rainfall in the Alps–Mediterranean Euroregion., J. Appl. Meteorol. Clim., 55, 655–671, doi:10.1175/JAMC-D-15-0095.1, 2016.

Scorzini, A. and Frank, E.: Flood damage curves: new insights from the 2010 flood in Veneto, Italy., J. Flood Risk Manag., –, 1–12, doi:10.1111/jfr3.12163, 2015.

Sebben, M., Werner, A., Liggett, J., Partington, D., and Simmons, C.: On the testing of fully integrated surface-subsurface hydrological models., Hydrol. Process., doi:10.1002/hyp.9630, 2012.

Smith, R.: The influence of mountains on the atmosphere., Adv. Geophys., 21, 87–230, doi:10.1016/S0065-2687(08)60262-9, 1979.

Todini, E.: A mass conservative and water storage consistent variable parameter Muskingum-Cunge approach., Hydrol. Earth Syst. Sc., 11, 1645–1659, doi:10.5194/hess-11-1645-2007, 2007.

Viero, D., D'Alpaos, A., Carniello, L., and Defina, A.: Mathematical modeling of flooding due to river bank failure., Adv. Water Resour., 59, 82–94, doi:10.1016/j.advwatres.2013.05.011, 2013.

Viero, D., Peruzzo, P., Carniello, L., and Defina, A.: Integrated mathematical modeling of hydrological and hydrodynamic response to rainfall events in rural lowland catchments., Water Resour. Res., 50, 5941–5957, doi:10.1002/2013WR014293, 2014.