Advancing global flood hazard simulations by improving comparability, benchmarking, and integration of global flood models

Environmental Research Letters

LETTER

Advancing global flood hazard simulations by improving comparability, benchmarking, and integration of global flood models

Jannis M Hoch[1,2] and Mark A Trigg[3]

[1] Department of Physical Geography, Faculty of Geosciences, Utrecht University, the Netherlands
[2] Unit Surface Water Systems, Deltares, Delft, the Netherlands
[3] School of Civil Engineering, Faculty of Engineering, University of Leeds, Leeds, United Kingdom

E-mail: j.m.hoch@uu.nl

Keywords: global flood models, model validation, framework, flood risk management, model intercomparison

OPEN ACCESS

RECEIVED
10 August 2018

REVISED
10 November 2018

ACCEPTED FOR PUBLICATION
26 November 2018

PUBLISHED
19 February 2019

Abstract

In recent years, a range of global flood models (GFMs) were developed, each utilizing different process descriptions as well as validation data sets and methods. To quantify the magnitude of these differences, studies assessed the performance of GFMs only on the continental and catchment level. Since the default model set-ups resulted in locally marked deviations, there is a clear need for further and especially more standardized research to not only maintain credibility, but also support the application of GFM products by end-users. Consequently, here we conceptually outline the basic requirements and challenges of a Global Flood Model Validation Framework for more standardized model validation and benchmarking. With the proposed framework we hope to encourage the much needed debate, research developments in this direction, and involvement of science with end-users. By means of the framework, it is possible to streamline the data sets used for input and validation as well as the validation approach itself. By subjecting GFMs to more thorough and standardized methods, we think their quality as well as acceptance will increase as a result, especially amongst end-users of their outputs. Otherwise GFMs may only serve a purely scientific purpose of continued 'siloed' model improvement but without practical use. Furthermore, we want to invite GFM developers to make their models more integratable which would allow for representation of more physical processes and even more detailed comparison on a model component basis. We think this is pivotal to not only improve the accuracy of model input data sets, but to focus on the core of each model, the process descriptions. Only if we know more about why GFMs deviate, are we able to improve them accordingly and develop a next generation of models, not only providing first-order estimates of flood extent but supporting the global disaster risk reduction community with more accurate and actionable information.

1. Introduction

Economic damage and casualties due to flooding increased remarkably in recent decades. Due to a combination of factors, such as population growth, urbanisation, and a changing climate, flood risk will continue to rise world-wide (Munich Re 2010, Ceola *et al* 2014, Winsemius *et al* 2016). For the implementation of improved flood risk management as well as efficient adaptation and mitigation measures a better understanding of the processes driving flood events is required. With most riverine flood events simultaneously impacting multiple neighbouring countries and catchments (Jongman *et al* 2014), declining availability of observed discharge data, and increasing computational power, the benefit of using global flood models (GFMs) was recognized as a key tool in tackling these challenges. Hence, their development and application increased rapidly in recent years (Ward *et al* 2015, Bates *et al* 2018).

All available GFMs are fit for the purpose of modelling global flood hazard and risk and validated to some extent during their development and dissemination. Yet, they all inherently have, depending on their

governing processes and structure, distinct properties, strengths, and weaknesses. Since validation data, period, and location are usually not consistent between GFM description studies, model differences do not directly become visible while in fact they can result in locally remarkable deviations when compared with each other (Trigg *et al* 2016, Bernhofen *et al* 2018).

In contrast to GFMs, other modelling groups are ahead in benchmarking model schemes and components. For instance, global hydrologic models and their components are regularly compared, such as their routing scheme (Zhao *et al* 2017) or simulated runoff (Beck *et al* 2017). Also, the land model community initiated a benchmarking project with regular meetings (Hoffman *et al* 2017) and the climate model community even developed a downloadable diagnostic and performance metrics tool for routine evaluation of models (Eyring *et al* 2016). Such model intercomparison projects are a great way to narrow the above-mentioned knowledge gap, let alone the stimulus for intensified scientific collaboration and exchange. To our knowledge, there is no such detailed comparison yet for GFMs besides first benchmarking efforts focussing on overall model only (Trigg *et al* 2016, Bernhofen *et al* 2018). Consequently, it is fair to say that GFMs are behind in terms of collaborative testing as they lack of more consistent and regular comparison, hampering a better understanding of the discrepancies in model outputs. This epistemic uncertainty could, we postulate, lead to problematic model equifinality as results may agree only coincidentally.

A better understanding of why and where each model may or should be used is, however, pivotal. Discerning not only a model's strengths, but also its uncertainties, limitations, and differences with respect to other models is a central pillar to put model outputs into perspective and increase their credibility. By virtue of a transparent comparison process with other models, individual model developers can see more clearly how to improve their own data sets and process representation where they may see these lacking.

Since the relative accuracy would become more tangible, the meaningfulness and applicability of each model for end-users would increase too. Various workshops aimed at bringing together researches and practitioners provide evidence that there is a growing demand for more transparency and better overview of GFMs as well as their characteristics and uncertainties (Salamon *et al* 2016, Salamon *et al* 2017, Willis Towers Watson 2018). This is particularly important for non-expert users of model outputs who rely on a clear understanding of the appropriateness and limitations in order to use the data appropriately (Ward *et al* 2015, Trigg *et al* 2016).

So, what are some possible ways forward? First, to facilitate obtaining the required understanding, an easily accessible yet demanding validation and benchmarking framework could create a meaningful starting point. This need is demonstrated by similar developments, for instance towards a framework for operational flood risk management (Alfieri *et al* 2018), as well as from the above-mentioned need of end-users to get a better grasp of model properties. Second, models are in almost all cases closed systems where output is produced based on the input provided and the subsequent model steps executed. While this works well for default model applications, it hampers the model's extension and integration with new features, components or even other models. In times of continued model integration, however, establishing links with other models via interfaces such as the application programming interface, open modelling interface or basic model interface (BMI) can facilitate including additional (physical and non-physical) processes simulated by other models.

In the remainder of this article, we first present a range of state-of-the art GFMs and outline their specific properties. Second, we assess the different validation data sets, periods, and locations of these GFMs as published in peer-reviewed papers to supplement our call for streamlined validation approaches. Subsequently, we sketch a possible design of a Global Flood Model Validation Framework for model validation and benchmarking. Last, motivation and possible approaches to advance the openness and integration capability of GFMs is presented. The article is ended with conclusions, ideas on how to implement the presented ideas and recommendations for further improvement of comparability and applicability.

2. Current GFMs

Currently, the openly accessible state-of-the-art GFMs are CaMa-Flood (Yamazaki *et al* 2011), GLOFRIS (Winsemius *et al* 2013), JRC (Dottori *et al* 2016), CIMA-UNEP (Rudari *et al* 2015), as well as the Fathom model (formerly SSBN; Sampson *et al* 2015) and the ECMWF model (Pappenberger *et al* 2012). These models can be divided into two main categories of GFMs, depending on the flow derivation modelling steps taken (figure 1).

It must be noted that there is also a number of private or national CAT models ('catastrophe models') that include global flood hazard, each also having its own properties, modelling cascades, and evaluation procedure and criteria. Obtaining information about these CAT models is, however, complicated due to the protection of intellectual property (IP) rights and competitive commercial advantage. The following comparison therefore represents only the most open models and may need updating in the future if these commercial models become more transparent.

The differing operations at various model stages result in a range of modelling approaches, each one

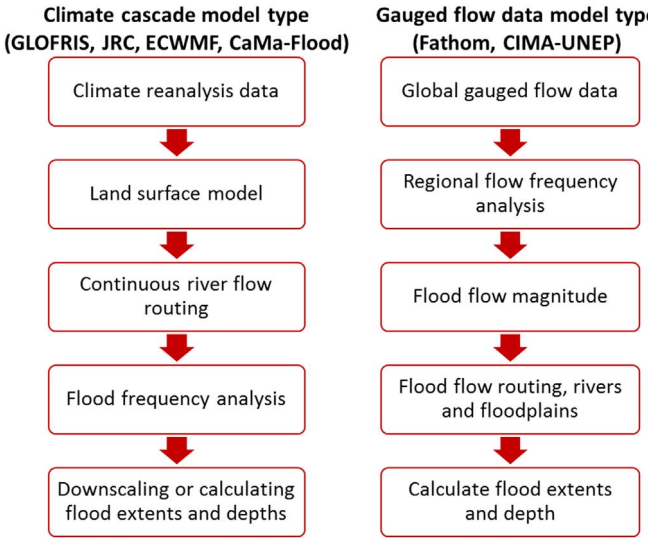

Figure 1. Modelling steps required for the two sub-categories of global flood models; modified from Trigg *et al* (2016).

using its own input data, method of calculating floodplain inundation, and spatial resolution. For instance, GLOFRIS runs at a 30 arcmin spatial resolution before post-processing and downscaling to 1 km, whereas the Fathom model yields output directly at 90 m globally. Such discrepancy in spatial resolution is possible because the models simulate processes with different scaling potential (Bierkens *et al* 2015). As a result, the models perform differently in these scale-dependent processes. For example, GFMs employing a land surface or hydrologic model excel in simulating the different water balance components. Contrariwise, the routing schemes of land surface or hydrologic models (typically the kinematic wave approximation) are less sophisticated than hydrodynamic models employing higher-order approximations of full shallow water equations. Therefore, their skill in simulating peak discharge (Zhao *et al* 2017) or backwater effects (Meade *et al* 1991), which are critical for flood hazard mapping, is curtailed.

Notwithstanding the differences, all models are applied regularly and used to inform policy-makers about flood hazard and risk. GLOFRIS, for example, is applied within the World Resources Institute Aqueduct tool, projecting current and future flood risk across the entire globe (World Resources Institute 2018). The JRC model is applied as part of the Global Flood Awareness System GloFAS (Alfieri *et al* 2013), and the Fathom model was recently used to compare flood risk with Federal Emergency Management Agency estimates across the continental United States (Wing *et al* 2018). CIMA-UNEP was applied for estimating current and future flood risk for the Global Assessment Report of the United Nations Office for Disaster Risk Reduction (Rudari *et al* 2015, UNISDR 2015).

3. META study: validation of GFMs

Before employing a GFM for flood hazard and risk assessments, ideally it should undergo thorough testing and validation. Due to the wide range of available observation data sets and depending on the model period as well as study area opted for, all GFMs may obtain good validation results, yet without providing any insight into performance relative to other GFMs. To get a grasp of the differences in model validation, here we detail the various data sets, periods, and locations used for the above-mentioned models.

As table 1 shows, the spread in validation (or benchmarking) data sets used is tremendous. Partially, this can be explained by the particular moment of model publication and the availability of data sets at that time. It also shows that most GFMs are validated against inundation extent, but only a few compare simulated discharge and water surface elevation with observations, although these aspects are important for flood risk management as well. Besides, the river basins, used for model validation, differ widely between studies as does the number of scientific reports documenting the model development over time.

Trigg *et al* (2016) showed that the GFMs listed in chapter 2 agree only for around a third of simulated flood extent in Africa. Since there can only be one actual realization of inundation at a time, this finding shows that a successful individual validation of models without comparison may be misleading with respect to the accuracy of the resulting inundation maps.

Together with the lack of congruency in model validation, the results from Trigg *et al* (2016) bolster the above-made claim that, to really get an idea of why model results deviate and to eventually learn from

Table 1. Overview of validation data sets of discharge, inundation extent, and water surface level (WSL) as used in various scientific studies of GFM development.

GFM	Study	River basin	Period	Validation data sets		
				Discharge	Extent	WSL
CaMa-Flood	Yamazaki *et al* (2011)	Amongst others: Amazon, Congo, Brahmaputra, Rhine, Ob	*Varying per basin*	GRDC[a]	SAR imagery (Hess *et al* 2003); GIEMS (Prigent *et al* 2007)	—
	Yamazaki *et al* (2012)	Amazon	2003–2005	ANEEL[b]	GIEMS (Prigent *et al* 2007)	Envisat RA-2
	Yamazaki *et al* (2013)	*Global maps used*	1991–2000	GRDC	—	—
	Yamazaki *et al* (2014)	Mekong	2001–2005	Inomata and Fukami (2008); MRC[c]	—	MRC[c]
GLOFRIS	Winsemius *et al* (2013)	Ganges–Brahmaputra	1961–1990	—	DFO[d]	—
JRC	Dottori *et al* (2016)	Tocantins, Severn, Thames, Elbe, Po, Niger, Indus, Ganges, Mekong, Irrawaddy	2000–2013		DFO; UNOSAT	—
		Amongst others: Rhine, Danube, Columbia, Thames, Colorado, Yukon		Local observations, based on (Hirpa *et al* 2016)		—
CIMA-UNEP[e]	Rudari *et al* (2015)	Colombia, Germany and Thailand	—	Amongst others: GRDC, RivDIS[f]	DFO	—
Fathom	Sampson *et al* (2015)[g]	Bow River, North Saskatchewan, Red Deer; Severn, Thames	*Comparing return periods*	—	Alberta State Government; JRC model	—
	Wing *et al* (2017)[g]	Conterminous United States	*Comparing return periods*	—	FEMA[h], USGS[i]	—
ECMWF	Pappenberger *et al* (2012)[g]	*Various major catchments on all continents*	*Comparing return periods*	—	Flood hazard maps as used by UNISDR	—

[a] GRDC, Global Runoff Data Centre.
[b] ANEEL, Agencia Nacional de Energia Electrica.
[c] MRC, Mekong River Commission.
[d] DFO, Dartmouth Flood Observatory.
[e] Development and validation of the CIMA-UNEP model was not published in peer-reviewed scientific journals.
[f] RivDIS, Global River Discharge Database.
[g] both studies only performed benchmarks with inundations maps from other inundation models or databases for given return periods.
[h] FEMA, Federal Emergency Management Agency.
[i] USGS, United States Geological Survey.

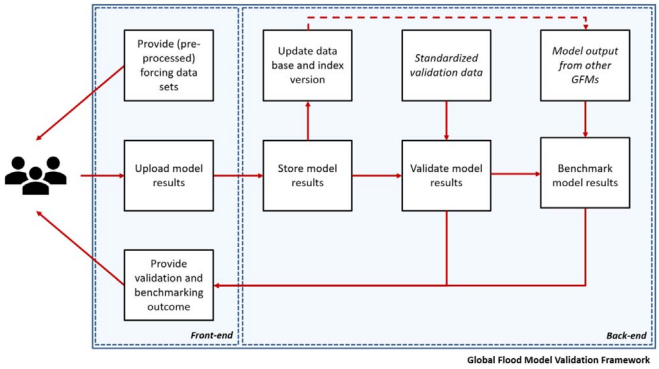

Figure 2. Conceptual design of a cloud-based Global Flood Model Validation Framework for model validation and benchmarking as well as maintaining a reference database.

each other, more standardized validation and benchmarking procedures could be useful.

4. Establishing a GFM validation framework

To facilitate standardized validation and benchmarking of models and their results, a framework facilitating these steps is needed. A first step towards more systematic benchmarking was set with the GLOFRIM framework by Hoch *et al* (2017) which allows for forcing different hydrodynamic models with identical hydrologic output. Yet, it can only mark a first proof-of-concept since much more functionality would eventually be needed. Some key tasks of a Global Flood Model Validation Framework would be, amongst others, to provide a front-end where users can upload model results as well as a back-end to not only execute validation and benchmarking autonomously but also store validation and benchmarking results (figure 2). Besides, the framework should provide input data sets to be used for each GFM run.

In its proposed conceptual form, the framework would be designed to only detect differences in simulated flood hazard. Since all GFMs employ different ways of how to determine risk by accounting for exposure and vulnerability, these aspects should be compared at a later stage as well. We here, however, focus on the physical modelling side of flood risk only to keep the scope of the study and proposed framework manageable. Moreover, many end-users such as insurance companies do have their own exposure and vulnerability maps and rely on hazard estimates for risk assessments.

By means of the framework, it would not only be possible to provide standardized input and validation data, but also to clearly define model boundary conditions. Using identical data will improve the comparability of model validation as this is currently done independently, using different validation data products, time periods, and study areas as shown in chapter 3.

4.1. Testing elements

We think it would be essential to test the models for the specific primary aspects listed below, yet this may be extended or altered if needed at any stage:

A. *Inundation extent*. A key output needed, for instance, by re-insurers to define flood-prone areas and determine premiums for portfolio exposure that intersects with the flood extent.

B. *Inundation depth*. Model output required by many risk assessments to assess potential damage via a depth-damage curve.

C. *Discharge hydrograph*. This is the fundamental driver of the out-of-bank flood processes.
By subjecting the GFMs to a thorough comparison and streamlining their input boundary conditions, the impact of the following secondary model aspects can also be tested:

D. *Forcing/Input data*. Assessing its impact is paramount to understand to which extent GFM accuracy is defined by model design or input/forcing data, something not covered by the study of Trigg *et al* (2016).

E. *Regionality*. Here referred to as a model's ability to perform in certain regions, differing in their meteorological, geographical, and other properties. Also, this could include an assessment whether GFMs perform well only for large rivers and where the threshold lies in the accurate representation of inundations along smaller reaches.

4.2. Testing challenges

Before establishing the framework, several decisions have to be made and existing challenges addressed, as Alfieri *et al* (2018) also pointed out. These decisions require but are not limited to the following list:

(a) *Test location*. First, is should be clear which river basins are ought to be used. As shown in table 1,

most major river basins were already used for validation and thus it would be sensible to use one of them. 'Classic' examples are the Amazon and Ganges–Brahmaputra basin as they both represent large low-lying floodplain areas where inundations occur regularly. The former is an indicator of performance in simulating large river flood extents while the latter is particularly relevant from a flood risk perspective due to a large population exposure and vulnerability. To be able to test for regionality, the chosen basins should also differ in their meteorological, geographical, and hydrological properties. Besides, only reasonably large catchments should be used to ensure that all models can sufficiently represent the processes despite differences in spatial resolution. As models and observations improve in resolution, the testing catchment testing scale can be adjusted appropriately.

(b) *Forcing data*. Despite most model forcing data being openly accessible, a clear decision has to be made which data set shall be used. For models based on meteorology (e.g. GLOFRIS or the JRC and ECMWF models), recent global forcing data such as ERA-Interim or ERA5 should be provided (Berrisford *et al* 2011, ECMWF 2018). In case derivatives such as flood wave hydrographs are required (e.g. Fathom, CIMA-UNEP), preprocessed and possibly bias-corrected data should be made available. In all cases, the data must be downloadable via the front-end of the framework.

(c) *Downstream boundaries*. Even though not all GFMs can accommodate dynamic sea levels as a downstream boundary condition, we recommend that in an initial approach this should be activated to facilitate comparison across default models settings. For advanced comparisons, the effect of changing downstream boundaries can be studied as well by de-activating them or, analogously, account for them once model development allows for it.

More challenging, validation data must be provided which meets the demands for state-of-the-art flood hazard modelling. State-of-the-art in this context also means that all validation data sets used must explicitly address possible uncertainties in observations. Hence, these additional aspects should be considered:

(d) *Discharge data.* Required to validate the models' skill to simulate discharge dynamics. Depending on the chosen test locations, different sources may be available, either global data sets or from local authorities. In case of the Amazon, for example, discharge data can be retrieved from ORE-HYBAM (ORE-HYBAM 2018). One of the few global databases of observed discharge is maintained by the Global Runoff Data Centre, currently containing data for around 1300 stations (GRDC 2018). To provide robust validation results, sufficiently long time series must be available. For those models simulating specific return periods only, corresponding discharge values should be derived from observations. One challenge may emerge from a possible mismatch between gauging station and river network. This issue was already solved for the CaMa-Flood model as described by Zhao *et al* (2017) and thus we recommend extending this approach to other GFMs.

(e) *Inundation maps.* Data ranging across various locations world-wide must be available, preferably open access remotely sensed satellite products to maintain global comparability. Since image quality may be hampered by cloud cover (Bernhofen *et al* 2018), this step may require some pre-processing. Alternatively, already pre-processed maps may be used, for example from the AquaMonitor (Donchyts *et al* 2016). Also, maps from the Dartmouth Flood Observatory (http://floodobservatory.colorado.edu/), the GIEMS data set (Prigent *et al* 2007, Papa *et al* 2010) or from Tellman *et al* (2017) can be used. For those models simulating inundation extent for specific return periods, inundation maps for actual events with corresponding return periods should be used as much as possible for validation, possibly building upon recent methods (Huang *et al* 2014, Giustarini *et al* 2015). Since properties of model and validation maps such as spatial or temporal resolution may deviate, matching model output to observations may be required.

(f) *Water levels.* To guarantee a globally uniform approach, satellite products should be used, for example ICESAT, ICESAT 2, ENVISAT or SWOT once available. The locations used for validation and benchmarking should be chosen such that potential vertical inaccuracies are limited. Again, the data must be carefully selected and preprocessed, for example to remove measurements affected by land or vegetation signals or differences due to different geoids used by model and satellite.

In a nutshell, the proposed framework's objectives are threefold: (1) provide forcing data, (2) validate and benchmark model results, and (3) store reference model output per GFM (figure 2). Once the user-performed simulation runs with the provided forcing data, results can be uploaded via a front-end to the framework's back-end. Here, both validation and benchmarking will be performed. For the validation, we suggest the following metrics: (i) for inundation extent, the hit ratio H, the false alarm ratio F, and the

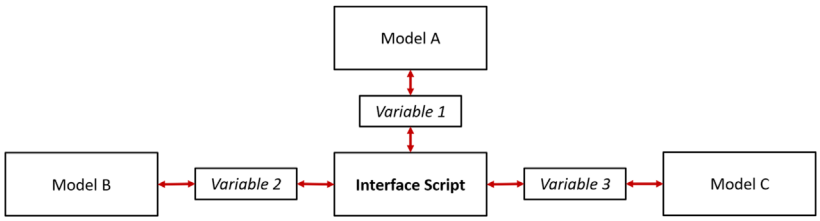

Figure 3. Schematic of coupling models as well as exchanging model information via BMI and a central interface script.

critical success index C; (ii) for discharge, the Kling-Gupta-Efficiency KGE (Gupta *et al* 2009) and its individual components; (iii) for surface water elevation, RMSE. Since these objective functions are only a recommendation, a definite choice should only be made after both developers and end-users agreed on common standards meeting their expectations and needs. This requires involvement of potential end-users in the development of the framework.

To perform the necessary operations, employing the increasing power of cloud computing could be a viable option. For benchmarking purposes, the model results will be stored in cloud-optimized format (for example cloud-optimized GeoTIFF; COG) and version-controlled according to the version number of tested GFM in a reference data repository, hence containing the most recent outputs of GFMs and allowing for tracking the impact of model developments on output. The reference observation data sets will then be used to apply the same objective functions. Once all steps are successfully executed, the resulting validation and benchmark statistics will be made available to the user via the front-end again.

The framework and data could be hosted by a neutral institution or other body, for instance within the Global Flood Partnership which already collected first experiences with a common tool for operational flood risk management (Alfieri *et al* 2018). Alternatively, such a framework could be hosted under the umbrella of the upcoming Global Risk Assessment Framework which aims at implementing a range of models and with a particularly end-user orientation (UNISDR 2018).

We are aware that setting up such a framework requires both financial and time resources. Yet, we believe that once validation and benchmarking of GFMs is streamlined, they will benefit by reducing uncertainty associated with model output and its application. Using centrally provided data would also enhance the reproducibility of model output, as work flows and data use would become more transparent. We are confident the efforts made will eventually pay off as model output uncertainty will be reduced while scientific discourse will be improved, leading to better informed decisions and reduced economic damage and casualties.

5. Opening the black box of model code

With the scientific funding bodies increasingly requiring research to be openly available, most (unfortunately not all) GFMs can be downloaded freely, advancing the usability and impact of the models. However, even with open code and model output availability, most models follow a 'black box' modelling approach of reading input data, executing a prescribed and model dependant set of processes, and thereafter providing output data (figure 1). Such approaches, nonetheless, pose a major limitation to making GFMs more integratable, intuitive, and interactive due to the lack of process accessibility. However, we consider process integration as key to better comparability as well as future improvements, and thus think that global flood hazard simulations can greatly benefit from opening the black box.

Admittedly, the integration of different models is not rocket science and was already achieved. Models can, for instance, use the output of model A as forcing for model B (Lian *et al* 2007, Biancamaria *et al* 2009, Schumann *et al* 2013). Such offline-coupling, however, increases overall computation time and may yield large intermediate files. Other forms of coupling entail online-coupling where the exchange of variables during model execution and without intermediate files is hard-coded (Viero *et al* 2014, Sutanudjaja *et al* 2017). Clearly, such bespoke model coupling is fit for bespoke purposes, yet it lacks the flexibility to be easily altered for other applications or to be extended with other models or only parts thereof.

To facilitate interactive and intuitive model coupling as well as to avoid 'integronster', i.e. models whose combined code is hard to disentangle and uncertainties are hard to trace (Voinov and Shugart 2013), the (BMI; Peckham *et al* (2013)) provides a powerful and flexible tool to exchange model information via an (user-defined) interface script (figure 3) without the need of integrating actual model code into one overarching model as exemplified with the EMELI framework (Peckham 2014).

The different models can for example be hydrologic or hydrodynamic models and exchange variables such as runoff or inundation depth, respectively. Yet, also other models could be linked up such as coastal or crop growth models or even non-physical models such as agent-based models.

Within the context of model benchmarking, implementing the BMI functionalities into GFMs may facilitate forcing them with identical data and, in turn, more standardized validation and benchmarking. In addition to the EMELI framework, the applicability of the BMI concept was shown by applying the GLOFRIM framework to benchmark different hydrodynamic models (Hoch *et al* 2017) as well as different schematizations of the same hydrodynamic model (Hoch *et al* 2018).

Unfortunately, none of the above-mentioned GFMs currently contains any BMI functionality (or anything like it). Since the implementation of a BMI is non-invasive, we think more efforts should be directed towards advancing the accessibility of model processes and variables by implementing this interface. In the long term, model integrability via BMI would, besides supporting model validation and benchmarking, allow for a plug-and-play design where applicants can create their 'own' GFM depending on their study specific needs and would also facilitate more efficient modelling efforts. That establishing such a framework is feasible was shown by the examples of EMELI and GLOFRIM.

Conveniently, the proposed framework can help in identifying which components of which models excel. For example, if benchmarking results indicate that Model A may profit from more physical groundwater modelling, such a module from another Model B could be added and forced with variables from Model A, for instance surface water depth. Besides, output from other non-GFM models could be employed such as sea levels from a tide and surge model (Model C) which would even further increase the number physical processes representable (figure 3).

We are aware that this would not only require opening the black box, but possibly also developers' minds. Besides, possible issues with IP rights may have to be solved first. Still, we are confident that such interactive model functionalities can become a core element of advancing model validation and benchmarking across scales and processes, as they may not only result in improved model skill but also new and promising research possibilities.

6. Conclusion and recommendations

Many GFMs have been successfully used in policy making as well as operational tools and systems, but a lack of model inter-comparison has led to a poor understanding of their differences. However, we think that discerning these differences is pivotal for increased acceptance of GFMs by end-users and thus the existing different approaches to simulate inundation data require a more thorough and streamlined validation and benchmarking procedure.

GFMs were validated with a wide range of data sets for various time periods in numerous river basins all over the World. While the data used for validation is to some extent related to the date of model publication and the data availability at that time, the fact that all models are validated 'successfully' for non-identical settings may lead to the misleading conclusion that all model perform equally well. Additionally, it does not support a clear conclusion as to why results differ between GFMs.

Due to the range in validation approaches, we see great potential for models to improve by comparing with and learning from others. Therefore, we sketch the conceptual outlines of a Global Flood Model Validation Framework serving multiple purposes. First, it provides identical model forcing. Second, it validates simulated discharge, inundation extent, and surface water elevations. Third, it serves as a repository and version-control of GFM output and thus also allows for benchmarking output from different models and model versions. By establishing such a framework, we can ensure that, despite all independent model development trajectories, the same data and criteria are applied for assessing model output. While we focus on the need for a framework, as well as key testing elements and challenges, we acknowledge that there will be multiple technical hurdles along the way. Even though we addressed several in this article, focussing on and solving all of them is, however, outside of the scope of this study.

Since the framework can only streamline external factors, there will probably still be deviations in model results due to differences in internal model structure, processes, and parameterization such as the use of different river networks between models. For now this is perfectly acceptable, as the proposed framework is not meant to converge all GFMs, but rather as a testing and learning environment for researchers to improve usability and acceptance by end-users. In the long run and after reducing uncertainty associated with model forcing and boundary conditions, the knowledge gained through the framework could help to disentangle the impact of internal model properties such as the spatial representativeness of model grid and elevation data and the use of different river networks.

By means of the framework, insights could be provided in the upsides and downsides of each tested model design. If the framework is applied for more river basins and hydrologic conditions, it would furthermore be possible to identify where and under which circumstances each model performs best. Such knowledge can, in turn, be beneficial when it comes to communicating model strengths and limitations to policy and decision makers and provides them with a tool to identify which GFM may be most appropriate for a project or application in a specific region. Besides, the insights gained may be used to better point towards model shortcomings which could benefit from adopting methods implemented in better

performing models. For models to profit further from such insights, it could be necessary to open the default 'black box' of model processes. While a standard comparison framework may be sufficient for default applications of the models, implementing functions to allow for accessing and exchanging model variables could facilitate integrating components from other GFMs to improve model performance. Moving away from a black box approach may further stimulate the benchmarking and comparison of GFMs, as assessments could be performed at an unprecedented level of detail and flexibility, allowing ranking of the importance of different elements of GFMs. For example, the same spatially varying hydrologic output could be applied to all models, reducing the number of factors influencing model deviations. Vice versa, it could be possible to provide a clearer picture on how the routines calculating hydrologic forcing may differ by applying one routing scheme to all models' designs. Ultimately, the GFMs would move closer together without abandoning their specific properties, and uncertainties surrounding flood hazard outputs could be reduced greatly.

We are aware that the presented conceptual framework and the required openness about model performance may discourage contributions from private CAT models. Nevertheless, we are convinced that an independent validation and benchmarking framework can be beneficial for the private sector too, as (a) data providers could present their results from commercial CAT models in a broader context, and (b) data users could first analyse which products fits their needs best before purchasing a flood product. We hope that thorough benchmarking of inundation maps becomes the new normal, eventually requiring vendors to improve their services and consequently resulting in better risk estimates for end-users. From a technical point of view implementing the CAT models into the proposed framework would be relatively straightforward as they essentially employ the same technology and input data types as the open scientific models. A major requirement for those model developers would of course be that outcomes are not necessarily made publicly available.

While the here proposed Global Flood Model Validation Framework focusses on differences in model design and associated differences in model output, more steps should be taken to improve the comparability and consequential uptake of GFMs. First, the nomenclature of model variables and components differs greatly between models, hampering the traceability of model work flows. By using more standardized terminology, for example the standard names proposed by the Community Surface Dynamics Modelling System (https://csdms.colorado.edu/wiki/CSDMS_Standard_Names), comparing GFMs would become easier, particularly for non-expert users. And second, comparability, inter-operability, and usability of model outputs would be greatly supported by agreeing on clear standards for files, for instance based on the guidelines of the Open Geospatial Consortium (http://opengeospatial.org/). Third, it is necessary that all GFMs (as for models in general) provide easily comprehensible description of how they work and what their outputs represent.

To establish a full comparison between GFMs, exposure and vulnerability data should be compared as well. Since these data layers are not based on a modelling cascades, the proposed Global Flood Model validation and benchmarking framework may not be the right means. Nevertheless, we think that further investigation is needed to better understand to which extent differences in simulated risk assessment outputs are dependent on hazard, exposure or vulnerability. Eventually, the three pillars of risk could be compared altogether. Such an extensive inter-comparison project would help greatly to advance the current state of GFMs and to identify new research possibilities.

More efforts should be taken to advance our understanding of GFMs and their differences. With our proposed validation and benchmarking framework together with greater model accessibility, we see great potential for future model developments as well as an increased number of GFM applications and hope that model comparison will play a more significant role in future flood hazard modelling studies.

Acknowledgments

We thank two anonymous reviewers for their critical and helpful remarks on a previous version of this manuscript.

ORCID iDs

Jannis M Hoch https://orcid.org/0000-0003-3570-6436
Mark A Trigg https://orcid.org/0000-0002-8412-9332

References

Alfieri L, Burek P, Dutra E, Krzeminski B, Muraro D, Thielen J and Pappenberger F 2013 GloFAS-global ensemble streamflow forecasting and flood early warning *Hydrol. Earth Syst. Sci.* **17** 1161–75

Alfieri L *et al* 2018 A global network for operational flood risk reduction *Environ. Sci. Policy* **84** 149–58

Bates P D, Neal J, Sampson C, Smith A and Trigg M 2018 Progress toward hyerresolution models of global flood hazard *Risk Modeling for Hazards and Disasters* ed G Michel (Amsterdam: Elsevier) ch 9 pp 211–32

Beck H E, van Dijk A I J M, de Roo A, Dutra E, Fink G, Orth R and Schellekens J 2017 Global evaluation of runoff from 10 state-of-the-art hydrological models *Hydrol. Earth Syst. Sci.* **21** 2881–903

Bernhofen M *et al* 2018 A first collective validation of global fluvial flood models for major floods in Nigeria and Mozambique *Environ. Res. Lett.* **13** 104007

Berrisford P, Dee D P, Poli P, Brugge R, Fielding K, Fuentes M, Kållberg P W, Kobayashi S, Uppala S and Simmons A 2011 The ERA-Interim archive Version 2.0 *ERA Report Series* ECMWF, Shinfield Park, Reading (www.ecmwf.int/node/8174)

Biancamaria S, Bates P D, Boone A and Mognard N M 2009 Large-scale coupled hydrologic and hydraulic modelling of the Ob river in Siberia *J. Hydrol.* **379** 136–50

Bierkens M F P *et al* 2015 Hyper-resolution global hydrological modelling: what is next?: 'Everywhere and locally relevant' *Hydrol. Process.* **29** 310–20

Ceola S, Laio F and Montanari A 2014 Satellite nighttime lights reveal increasing human exposure to floods worldwide *Geophys. Res. Lett.* **41** 7184–90

Donchyts G, Baart F, Winsemius H, Gorelick N, Kwadijk J and van de Giesen N 2016 Earth's surface water change over the past 30 years *Nat. Clim. Change* **6** 810

Dottori F, Salamon P, Bianchi A, Alfieri L, Hirpa F and Feyen L 2016 Development and evaluation of a framework for global flood hazard mapping *Adv. Water Resour.* **94** 87–102

Copernicus Climate Change Service (C3S) 2017 ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate. Copernicus Climate Change Service Climate Data Store (CDS) (https://cds.climate.copernicus.eu/cdsapp#!/home) (Accessed: 21 December 2018)

Eyring V *et al* 2016 ESMValTool (v1.0)—a community diagnostic and performance metrics tool for routine evaluation of Earth system models in CMIP *Geosci. Model Dev.* **9** 1747–802

Giustarini L, Chini M, Hostache R, Pappenberger F and Matgen P 2015 Flood hazard mapping combining hydrodynamic modeling and multi annual remote sensing data *Remote Sens.* **7** 14200–26

GRDC 2018 GRDC Composite Runoff Fields v1.0 (http://grdc.sr.unh.edu/) (Accessed: 9 March 2018)

Gupta H V, Kling H, Yilmaz K K and Martinez G F 2009 Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling *J. Hydrol.* **377** 80–91

Hess L L, Melack J M, Novo E M L M, Barbosa C C F and Gastil M 2003 Dual-season mapping of wetland inundation and vegetation for the central Amazon basin *Remote Sens. Environ.* **87** 404–28

Hirpa F A, Salamon P, Alfieri L, Pozo J T, Zsoter E and Pappenberger F 2016 The effect of reference climatology on global flood forecasting *J. Hydrometeorol.* **17** 1131–45

Hoch J M, Neal J C, Baart F, van Beek R, Winsemius H C, Bates P D and Bierkens M F P 2017 GLOFRIM v1.0—a globally applicable computational framework for integrated hydrological-hydrodynamic modelling *Geosci. Model Dev.* **10** 3913–29

Hoch J M, Van Beek L P H, Winsemius H C and Bierkens M F P 2018 Benchmarking flexible meshes and regular grids for large-scale fluvial inundation modelling *Adv. Water Resour.* **121C** 350–60

Hoffman F M *et al* 2017 2016 international land model benchmarking (ILMB) workshop report *DOE/SC-0186* U.S. Department of Energy (https://science.energy.gov/~/media/ber/pdf/workshop%20reports/2016_ILAMB_Report.pdf)

Huang C, Chen Y and Wu J 2014 Mapping spatio-temporal flood inundation dynamics at large river basin scale using time-series flow data and MODIS imagery *Int. J. Appl. Earth Obs. Geoinf.* **26** 350–62

Inomata H and Fukami K 2008 Restoration of historical hydrological data of Tonle Sap Lake and its surrounding areas *Hydrol. Process.* **22** 1337–50

Jongman B, Hochrainer-Stigler S, Feyen L, Aerts J C J H, Mechler R, Botzen W J W, Bouwer L M, Pflug G, Rojas R and Ward P J 2014 Increasing stress on disaster-risk finance due to large floods *Nat. Clim. Change* **4** 1–5

Lian Y, Chan I-C, Singh J, Demissie M, Knapp V and Xie H 2007 Coupling of hydrologic and hydraulic models for the Illinois River Basin *J. Hydrol.* **344** 210–22

Meade R H, Rayol J M, Conceicão S C and Natividade J R G 1991 Backwater effects in the Amazon River basin of Brazil *Environ. Geol. Water Sci.* **18** 105

Munich Re 2010 Topics geo natural catastrophes 2009: analyses, assessments, positions 302-06295 Munich Reinsurance Group, Munich, Germany

ORE-HYBAM 2018 HYBAM ORE-South America (http://ore-hybam.org/index.php/eng/Data/Station-Access-Maps/HYBAM-ORE-South-America) (Accessed: 19 July 2018)

Papa F, Prigent C, Aires F, Jimenez C, Rossow W B and Matthews E 2010 Interannual variability of surface water extent at the global scale, 1993–2004 *J. Geophys. Res. Atmos.* **115** 1–17

Pappenberger F, Dutra E, Wetterhall F and Cloke H L 2012 Deriving global flood hazard maps of fluvial floods through a physical model cascade *Hydrol. Earth Syst. Sci.* **16** 4143–56

Peckham S D 2014 MELI 1.0: An Experimental Smart Modeling Framework For Automatic Coupling Of Self-Describing Models 9 CUNY Academic Works (https://academicworks.cuny.edu/cc_conf_hic/464)

Peckham S D, Hutton E W H and Norris B 2013 A component-based approach to integrated modeling in the geosciences: the design of CSDMS *Comput. Geosci.* **53** 3–12

Prigent C, Papa F, Aires F, Rossow W B and Matthews E 2007 Global inundation dynamics inferred from multiple satellite observations, 1993–2000 *J. Geophys. Res.* **112** D12107

Rudari R, Silvestro F, Campo L, Rebora N, Boni G and Herold C 2015 Improvement of the global flood model for theGAR2015 (https://preventionweb.net/english/hyogo/gar/2015/en/bgdocs/risk-section/CIMA%20Foundation,%20Improvement%20of%20the%20Global%20Flood%20Model%20for%20the%20GAR15.pdf)

Salamon P *et al* 2016 The Global Flood Partnership Conference *Linking global flood information with local needs (Ispra, Italy, 29 June–1 July 2016)* (Luxembourg: Publications Office of the European Union) (https://doi.org/10.2788/84223)

Salamon P *et al* 2017 The Global Flood Partnership Conference *From hazards to impacts (Ispra, Italy , 27–29 June 2017)* (Luxembourg: Publications Office of the European Union) (https://doi.org/10.2760/68734)

Sampson C C, Smith A M, Bates P D, Neal J C, Alfieri L and Freer J E 2015 A high-resolution global flood hazard model *Water Resour. Res.* **51** 7358–81

Schumann G J-P, Neal J C, Voisin N, Andreadis K M, Pappenberger F, Phanthuwongpakdee N, Hall A C and Bates P D 2013 A first large-scale flood inundation forecasting model *Water Resour. Res.* **49** 6248–57

Sutanudjaja E H *et al* 2017 PCR-GLOBWB 2.0: a 5 arc-minute global hydrological and water resources model *Geosci. Model Dev. Discuss.* **11** 2429–53

Tellman B, Sullivan J, Doyle C, Kettner A, Brakenridge G R, Erickson T and Slayback D A 2017 A global geospatial database of 5000 + historic flood event extents *AGU Fall Meeting Abstracts*

Trigg M A *et al* 2016 The credibility challenge for global fluvial flood risk analysis *Environ. Res. Lett.* **11** 094014

UNISDR 2015 Global Assessment Report on Disaster Risk Reduction, Making Development Sustainable: The Future of Disaster Risk Management, Geneve

UNISDR 2018 Putting science to work for resilience (https://unisdr.org/archive/58772)

Viero D P, Peruzzo P, Carniello L and Defina A 2014 Integrated mathematical modeling of hydrological and hydrodynamic response to rainfall events in rural lowland catchments *Water Resour. Res.* **50** 5941–57

Voinov A and Shugart H H 2013 'Integronsters', integral and integrated modeling *Environ. Modelling Softw.* **39** 149–58

Ward P J *et al* 2015 Usefulness and limitations of global flood risk models *Nat. Clim. Change* **5** 712–5

Willis Towers Watson 2018 Insights from the Willis Re Flood Club: the Weaknesses and Strengths of Flood Modelling (Willis Towers Watson Wire) (https://blog.willis.com/2018/06/insights-from-the-willis-re-flood-club-the-weaknesses-andstrengths-of-flood-modelling) (Accessed: 19 July 2018)

Wing O E J, Bates P D, Sampson C C, Smith A M, Johnson K A and Erickson T A 2017 Validation of a 30 m resolution flood hazard model of the conterminous United States *Water Resour. Res.* **53** 7968–86

Wing O E J, Bates P D, Smith A M, Sampson C C, Johnson K A, Fargione J and Morefield P 2018 Estimates of present and future flood risk in the conterminous United States *Environ. Res. Lett.* **13** 034023

Winsemius H C, van Beek L P H, Jongman B, Ward P J and Bouwman A 2013 A framework for global river flood risk assessments *Hydrol. Earth Syst. Sci.* **17** 1871–92

Winsemius H C *et al* 2016 Global drivers of future river flood risk *Nat. Clim. Change* **6** 381–5

World Resources Institute 2018 Aqueduct Global Flood Analyzer (http://floods.wri.org) (Accessed: 8 March 2018)

Yamazaki D, Kanae S, Kim H and Oki T 2011 A physically based description of floodplain inundation dynamics in a global river routing model *Water Resour. Res.* **47** 1–21

Yamazaki D, Lee H, Alsdorf D E, Dutra E, Kim H, Kanae S and Oki T 2012 Analysis of the water level dynamics simulated by a global river model: a case study in the Amazon River *Water Resour. Res.* **48** 1–15

Yamazaki D, De Almeida G A M and Bates P D 2013 Improving computational efficiency in global river models by implementing the local inertial flow equation and a vector-based river network map *Water Resour. Res.* **49** 7221–35

Yamazaki D, Sato T, Kanae S, Hirabayashi Y and Bates P D 2014 Regional flood dynamics in a bifurcating mega delta simulated in a global river model *Geophys. Res. Lett.* **41** 3127–35

Zhao F *et al* 2017 The critical role of the routing scheme in simulating peak river discharge in global hydrological models *Environ. Res. Lett.* **12** 075003