

A quality-control framework for sub-daily flow and level data for hydrological modelling in Great Britain

Felipe Fileni ^{a,*}, Hayley J. Fowler ^a, Elizabeth Lewis ^a, Fiona McLay^b and Longzhi Yang ^c

^a School of Civil Engineering and Geosciences, Newcastle University, Newcastle upon Tyne NE1 7RU, United Kingdom

^b Scottish Environment Protection Agency (SEPA), Angus Smith Building, 6 Parklands Avenue, Eurocentral, Holytown, North Lanarkshire ML1 4WQ, United Kingdom

^c Department of Computer and Information Sciences, Northumbria University, Newcastle Upon Tyne NE1 8ST, United Kingdom

*Corresponding author. E-mail: F.De-Mendonca-Fileni2@newcastle.ac.uk

FF, 0000-0001-7260-1946; HJF, 0000-0001-8848-3606; EL, 0000-0002-7471-9988; LY, 0000-0003-2115-4909

ABSTRACT

The absence of an accessible and quality-assured national flow dataset is a limiting factor in sub-daily hydrological modelling in Great Britain. The recent development of measuring authority APIs and projects such as the Floods and Droughts Research Infrastructure (FDRI) programme aim to facilitate access to such data. Basic quality-control (QC) of 15-minute data is performed by the data collection authorities and the National River Flow Archive (NRFA). Still, there is a need for a comprehensible and verifiable quality control methodology. This paper presents an initial assessment of the available data and examines what needs to be done for applicability of the data at national scale. The 15-minute flow series has many inconsistencies, and there are also inconsistencies with the NRFA Annual Maximum values. When producing a QCed dataset, decisions regarding the retention of data values need to be taken and recorded. Furthermore, QC should remove and rectify erroneous values, such as negative and above world record flows; and an assessment of homogeneity and truncated values in the stations could be beneficial to flag suspect data. The complex chain for production and changeability of flow and level data makes data curation and governance imperative to assure the longevity of the dataset.

Key words: flow, large sample hydrology, level, observations, quality control, sub-daily

HIGHLIGHTS

- Sub-daily flow and level datasets are a step towards cutting-edge hydrological modelling in the UK.
- Currently available data have no consistent quality-control checks at a national level.
- Currently available data does not have traceable data versions.
- There are a lot of inconsistencies in the available dataset.
- A framework aiming to tackle these problems and produce a national product is presented.

INTRODUCTION

It was no surprise that the poll held during the Floods and Droughts Research Infrastructure (FDRI) panel discussion at BHS2022 pointed out that most hydrologists in the room identified improving the quality of hydrological data as the highest priority among hydrologists in the UK. With the turn of the century, hydrological records became longer and more available. With more data, a shift in hydrology has occurred from a demand for more observational data and a 'value of data' approach (Beven & Binley 1992; Beven 1993) to a heavier focus on the uncertainties in the input data before model fitting (Beven 2006). Currently, data quality is considered one of the key factors for the improvement of hydrological predictions and modelling (Beven 2019; Blöschl *et al.* 2019; Wagener *et al.* 2021). However, data measurements are surrounded by epistemic uncertainties (Beven 2021) that are often propagated or extrapolated in the generation of datasets that are suitable inputs for hydrological modelling (McMillan *et al.* 2012; McMillan *et al.* 2018, 2022).

The digital revolution has facilitated data sharing and has allowed the use of computer-intensive techniques in science. Paradoxically, it might have made science less reproducible. Components like data, models, code, and instructions must be made

This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence (CC BY 4.0), which permits copying, adaptation and redistribution, provided the original work is properly cited (<http://creativecommons.org/licenses/by/4.0/>).

accessible to fully reproduce a hydrological experiment. Although authors and institutions strive to enhance reproducibility, their efforts lag behind the rapid data revolution. While Journal policies such as data and code availability upon request improve experiment replicability, they remain insufficient for the majority of cases (Stagge *et al.* 2019). Trust in experiment reproducibility is dwindling; a survey noted that 90% of scientists believe that there is a reproducibility crisis in academia (Baker 2016). This extends to hydrological sciences where renowned hydrological journals host only 0.6–6.8% of fully reproducible articles, a key issue being the availability of the data used (Stagge *et al.* 2019).

Assuring datasets to have a certain degree of quality while making them accessible to the public is a challenge in hydrology. Manually checking large datasets is a time-consuming procedure that is subject to human errors. Hence, together with manual checks at a local level, done by station operators, performing automatic quality-control checks is necessary before making a dataset available. These checks can be very simple, like checking if gauged flow or rainfall is above a theoretically impossible threshold or avoiding physically impossible values (Coxon *et al.* 2015; Gudmundsson *et al.* 2018; Lewis *et al.* 2018b; Crochemore *et al.* 2020; Lewis *et al.* 2021), to more complex checks such as checking if rainfall gauges are consistent with nearby gauges (Lewis *et al.* 2021).

Another outcome of the digital revolution has been an incremental increase in the temporal and spatial resolution of data, improving the knowledge of physical processes in hydrology. Notably, the use of sub-daily rainfall observations has improved our understanding of short-duration rainfall extremes and their increasing intensity with global warming (Barbero *et al.* 2017; Prein *et al.* 2017; Guerreiro *et al.* 2018; Fowler *et al.* 2021). However, this increase in rainfall extremes does not necessarily translate to an increase in flood hazards (Hettiarachchi *et al.* 2019; Sharma *et al.* 2021), as there are multiple complex drivers of floods such as snowmelt and antecedent soil moisture conditions (Massari *et al.* 2014; Arheimer & Lindström 2015; Wasko & Nathan 2019). In smaller/flashy catchments, evidence of flood hazard increases due to climate change being more pronounced, as other drivers play a less important role in their physical processes (Wasko & Sharma 2017; Wasko & Nathan 2019). Despite this, one of the key issues in flood modelling in smaller catchments is the absence of reliable sub-daily resolution flow or level data. This is because the daily flood peaks tend to underestimate the true flood peak and potentially rapid rates of rise (Archer & Fowler 2021), even when the data are disaggregated into smaller time-steps, especially in smaller catchments (Chen *et al.* 2017; Beylich *et al.* 2021).

Using sub-daily resolution data for flood modelling could deliver a step change in identifying future flood events and mitigating flood damage. The UK already has a range of sub-daily rainfall datasets available, such as INTENSE (Gauge data – Blenkinsop *et al.* 2018) and CEH-GEAR (Gridded data – Tanguy *et al.* 2021), that are quality-controlled (Blenkinsop *et al.* 2017; Lewis *et al.* 2018b) and open-source. Historically, the use of sub-daily flow data has been limited to flood estimation methodologies in the Flood Estimation Handbook (FEH), which rely primarily on annual maxima (AMAX) extracted from a sub-daily flow time series. The sub-daily records are available upon request to the relevant authority, with the data being quality-controlled at a local level. Nevertheless, the data are dispersed and there are no consistency or quality-control checks performed at a national level, or traceable data versions. Such procedures could be highly beneficial for improving the quality of flow and level of data in the UK and are complementary to projects aiming to make data in the UK more accessible, such as Floods and Droughts Research Infrastructure (FDRI – <https://www.ceh.ac.uk/our-science/projects/floods-and-droughts-research-infrastructure-fdri>).

There is an evident shortage of sub-hourly flow/level data publicly/easily available in the UK, which will further be presented in the next section; then, some issues encountered with the datasets will be assessed; next, the reasons for these inconsistencies will be explained; finally, the article will discuss the challenges that need to be addressed to use the 15-min flow data of UK for hydrological modelling.

CURRENTLY AVAILABLE UK DATA

The UK has a wide variety of rainfall and flow products (Figure 1 and Supplementary material, Annex 1). Sub-hourly rainfall datasets are available in several formats, i.e., gauging, gridded and future projections, with potential to be used in a range of hydrological models. Nevertheless, their use, at a national scale, is limited, in part, by the availability of quality-assured streamflow datasets to validate these models. Continuous hydrological models at a national scale, such as Grid-to-Grid (Bell *et al.* 2009), FUSE-GB (Lane *et al.* 2019), LSTM-GB (Lees *et al.* 2021), and SHETRAN-GB (Lewis *et al.* 2018a) are set up at a daily temporal resolution, because quality-assured flow datasets, such as CAMELS-GB (Coxon *et al.* 2020) and the National River Flow Archive (NRFA), are only available at daily time-steps.

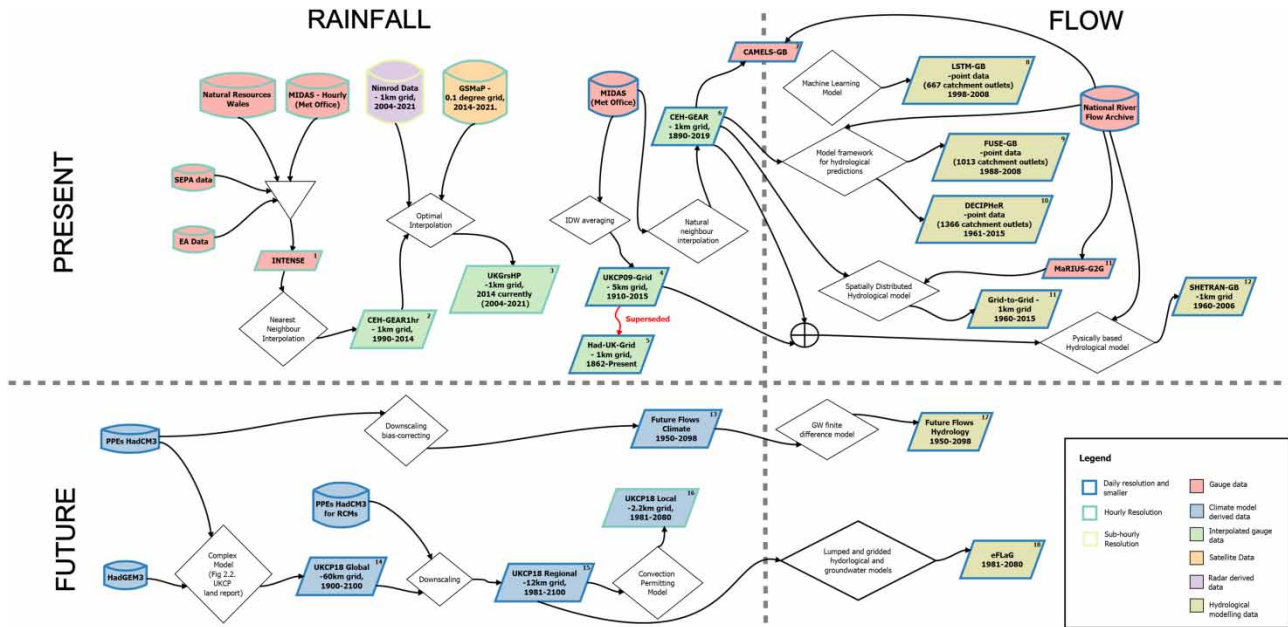


Figure 1 | Summary of daily and sub-daily rainfall and flow datasets available in the UK – for more information on a particular dataset please look at the superscript number and the reference in Supplementary material, Annex 1.

Similarly, statistical and hydrological models are utilized for the generation of future flow datasets, e.g., Future-Flows and e-Flag (Figure 1- Haxton *et al.* 2012; Hannaford *et al.* 2022). These will subsequently be used in predictive applications, such as defining climate change flow allowances (Kay 2021) and identifying potential trends in flows with climate change (Collet *et al.* 2018). The new UK future climate projections, UKCP18, have been made available at sub-daily temporal resolution, with new convection-permitting climate models applied to downscale regional to local projections – UKCP Local (Fosser *et al.* 2019). Still, sub-daily future flow predictions are limited by the calibration of hydrological models, and consequentially, the lack of a sub-daily flow database. Thus, even with the enhanced resolution offered by the UKCP Local, future flow models are still being run on a daily timescale, given the calibration data limitation.

In Great Britain, 15-min flow time series is recorded by the Environment Agency (EA-England), the Scottish Environment Protection Agency (SEPA-Scotland), and Natural Resources Wales (NRW-Wales). In a first step, the 15-min data are quality controlled, periodically, at the agency level. Then, the data are transformed into an AMAX and a peak-over-threshold (POT) dataset by the NRFA. These are the series used in industry-standard methodologies for flood predictions in the UK, with a further effort to quality control these peaks, manually, by the agencies and UKCEH. The procedure entails an annual check of a subset of the whole data available in the archive. Finally, an additional quality assessment is conducted to determine the station’s suitability to different analysis. A gauge is labelled as appropriate for QMED if the measurement error for QMED values does not exceed 30%. For stations considered for pooling, measurements of AMAX1, AMAX2, and AMAX3– the top three annual values – are considered. If these values are deemed accurate, that is, having errors for AMAX2 and AMAX3 below 30%, and AMAX1 being precise enough, the data are designated as suitable for pooling. Despite being semi-qualitative with no level of confidence associated with it (Wallingford HydroSolutions 2016), these checks play a crucial role in UK flood design analysis. They enable the categorization of stations based on their confidence levels in higher flows: differentiating between highly confident pooling stations, stations with a reasonable degree of confidence in higher flows suitable for QMED analysis, and stations providing inaccurate data for flows surpassing QMED.

For continuous data analysis, the 15-min records are available upon station request done to the agencies. Nevertheless, it’s important to note that the quality control and assurance measures implemented for peak flow values within the NRFA archive do not extend to the 15-min station time series. This discrepancy arises due to the focus of NRFA’s quality-control efforts exclusively on peak flow values, rendering them inapplicable to the continuous time series data.

Recently, access to the continuous data have recently been facilitated by the development of APIs, such as the SEPA, NRW, and EA (NRW 2016; SEPA 2022; Environment Agency 2023). However, these APIs are still subject to specific data

limitations. For instance, the EA API lacks the inclusion of all stations available in the NRFA annual maximum archive. The SEPA API imposes an initial restriction on the volume of data that can be downloaded within a single day. Also, the NRW API exclusively provides access to the most recent year's 15-min data.

METHODS

Data used

This study uses 15-min flow and level data from the UK agencies, and the NRFA AMAX flow and level series.

The SEPA 15-min data were downloaded from their APIs, with a provided access key giving access to a larger number of daily downloads. With the API, all the available gauging stations were downloaded, totalling 315 flow and 390 level stations, with 274/273 level/flow stations identified in the NRFA archive. NRW and the EA have provided data upon request, with raw datasets coming from WISKI, a software for hydrologic data storage. Due to measuring authority time and workforce constraints, only peak flow stations, both suitable for pooling and QMED, from the EA and NRW were requested: 607/74 flow and 608/76 level stations were identified from the EA/NRW provided downloads. The data obtained were the continuous/semi-continuous flow and level time series for these stations, alongside the quality-control code given by them at the agency level. Finally, the latest version of the NRFA AMAX series was downloaded from their website.

Pre-treatment of data

Before use, the 15-min time series from the agencies was standardized and joined. Often, local agencies store data with different headers and nomenclature, e.g., some EA agencies use different columns for date and time, while others use a date–time column. Another occurrence, in very large time series, were splits of data into different .csv files that often contained repeated dates. Finally, the time series of flow and level was capped until the year 2021, the last available full year of data; and to two decimal cases, for standardisation and data storage purposes.

Mismatch analysis

The next step was focusing on identifying and understanding the reason why mismatches in the 15-min time series occur. A mismatch is categorized as when a time series has duplicate date–time values with different flows or level values. Two types of mismatches were analyzed: *i* – duplicate dates with different values in the 15-min time series; *ii* – an analysis of AMAX values that did not match the NRFA AMAX values.

RESULTS

Data summary by country and date

Table 1 summarizes the stations available in the level and flow datasets. There is a noticeable difference among Scottish (SEPA) and English/Welsh (EA/NRW) data. The first was sourced directly from their APIs, and notwithstanding the shorter median length of the dataset, the Scottish data had fewer gaps and fewer stations with >10% data missing. There were also no mismatches within the Scottish time series. English and Welsh data, gathered upon request to the EA and NRW, presented similar statistics in terms of median length, gaps, and percentage of stations with >10% missing data (Table 1).

The UK has digital 15-min data tracing back to the 1930s (two level stations and one flow station), with an exponential growth of station coverage from the 1950s to the 1990s. Overall, hydrological flow and level data becomes more consistent in the 1980s, when spatial coverage reaches 80% of its current capacity; and most of the dataset does not have any temporal gaps in it (Table 2). A further increase in data completeness is perceptible from the 1980s to the 1990s, with stations going from an average of 83 (level) and 82 (flow)% of temporal coverage data to 94%. From the 2000s, the data reaches the spatial and temporal completeness levels as of today, with more than 90% of the stations with no value gaps within the decade. There is no definitive answer on why the improvements in the number of stations record and completeness happened, but some important changes in UK hydrometry and hydrology have happened in the end of last century: the release of the flood study reports, in 1975, and subsequent investments in gauging stations; the transition from charted data to digital records, that occurred mostly in the 1980s; an increased focus, from agencies, on accurately recording high flows for flood studies, rather than only low flows for water quality are all pointers of an increase in interest in the upkeep of flow and level records.

Table 1 | Summary of stations and mismatches in the 15-min flow and level dataset for the countries of Great Britain





Station type								
	Level	Flow	Level	Flow	Level	Flow	Level	Flow
Summary								
Number of stations	390	315	608	607	76	74	1,074	996
Number of stations NRFA	274	263	608	607	76	74	958	944
Years of data (total)	11,636	9,903	23,074	22,827	2,855	2,655	37,565	35,385
Median length (years)	30.86	31.12	40.01	39.87	39.15	38.89	39.54	38.25
Total gaps (years)	207.86	161.47	2,040.39	1,901.73	212.35	172.55	2,460.60	2,235.75
Number (%) of stations with gaps	323 (83)	281 (89)	587 (97)	580 (96)	69 (91)	68 (92)	979 (91)	929 (93)
Number (%) of stations with 10%+ gaps	23 (6)	14 (4)	140 (23)	130 (21)	12 (16)	14 (19)	175 (16)	158 (16)

Table 2 | Decadal summary and completion rate (mean, median completion, and stations with full data) of the 15-min flow and level dataset (H = level data, Q = flow data)

Decade Station type	50-59		60-69		70-79		80-89		90-99		00-09		10-19		20-21	
	H	Q	H	Q	H	Q	H	Q	H	Q	H	Q	H	Q	H	Q
Stations with some data	28	18	167	143	480	450	732	705	877	866	919	915	916	910	904	891
Mean completion	0.42	0.32	0.48	0.44	0.61	0.58	0.83	0.82	0.94	0.94	0.97	0.97	0.99	0.99	0.99	0.99
Median completion	0.35	0.29	0.38	0.34	0.71	0.62	1	1	1	1	1	1	1	1	1	1
Stations with full data	2	1	28	18	164	140	459	438	712	697	859	845	891	881	852	838

Mismatches

SEPA data were obtained directly from their APIs, as a continuous time series, without duplicates. In contrast, the other two agencies had duplicate records in their datasets. The 15-min data from Welsh and English measuring authorities are categorized into two types of files: an irregular (EA) or 0-s (NRW) time series, covering the earlier period until 2003/04 when the current storage system was implemented. These series, extracted from charted data, occasionally have flows with higher resolution than 15 min, especially during high peaks. However, there is no consistent pattern for these occurrences; they vary from station to station.

The second part of the dataset comprises systematically regular 15-min data, extending to the present day. Since the resolution of both time series differs – one maintains a constant 15-min resolution, while the other generally is recorded at 15-min intervals, but with irregular time-steps during peaks – they have not been merged and are stored separately. The issue of duplicate records arises due to temporal overlaps between the 15-min data and the irregular time series. This inconsistency becomes problematic when the same flow or level series do not present the same value at the date time-step.

Among the 681 stations in the EA and NRW dataset, 556 of them had at least one occurrence of duplicate date-times. In most cases, these duplicates contained identical information. However, there were 143 stations with the same date-time indexes that held different flow records. Most mismatches took place between the 1970s and the 2000s (Figure 2), with a progressive decrease in the percentage of mismatched stations from the 1970s to the late 1990s (Figure 2). A significant increase in mismatches is observed in the early 2000s, mainly in 2003, coinciding with the transition of data to the new system of storage, WISKIS. After 2004, only three stations presented overlaps.

Furthermore, it is worth noting that the 15-min data from all three agencies may occasionally not align with the NRFA AMAX dataset, even though they were extracted from the 15-min time series. A comparison between the 15-min data with version 10 of the NRFA AMAX archive shows that approximately 13% of the values between the time series and the archive

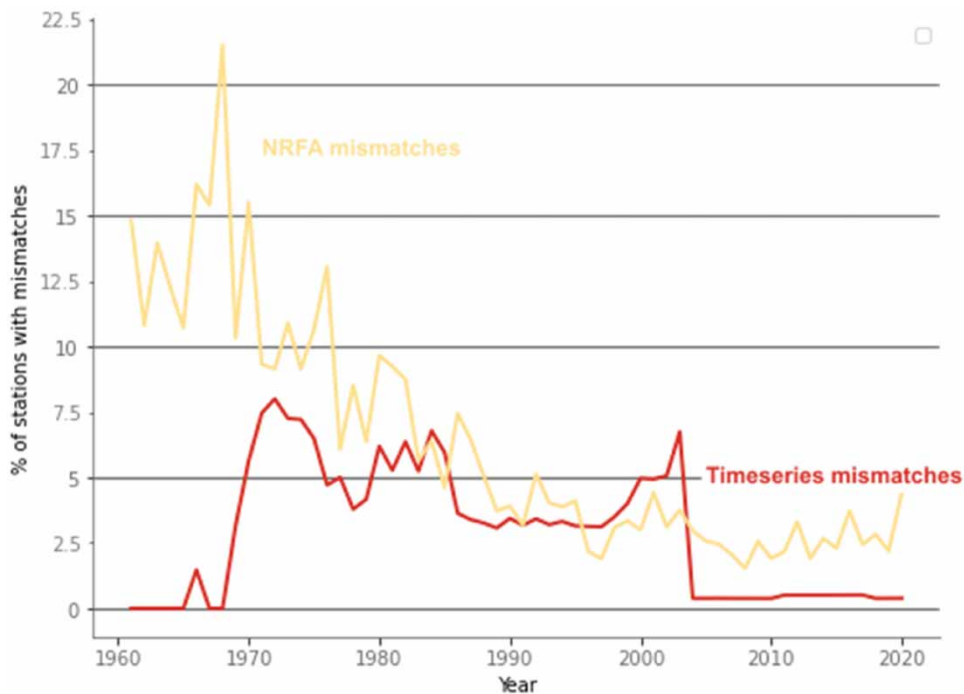


Figure 2 | Percentage of stations where mismatches (NRFA and within time series) occur per year.

differ. The occurrence of mismatches between these datasets is inversely related to the age of the record, with newer records demonstrating better agreement between both datasets (Figure 2).

DISCUSSION

Why do we need a quality-assured sub-daily flow and level dataset at a national scale?

In the UK, 15-min flow data have wide-ranging use in the calibration of industry-standard models. The FEH statistical method uses 15-min data to derive their POT series (Robson & Reed 1999). This is a necessity, as more than 75% of NRFA catchments have a time to peak (T_p) smaller than 8.25 h (Kjeldsen 2007). On these occasions ($T_p < 24$ h), daily intervals are insufficient to capture instantaneous peak flows happening during a flood in a catchment.

UK sub-daily flow time series has also seen some applicability in academia. For example, Prosdocimi *et al.* (2015) used sub-daily flow time series to investigate the effects of urbanization in extreme floods, by comparing similar catchments that only differed in their urbanization level. Hourly flow data for the River Axe hourly was used to identify how agricultural land use change could impact runoff in the region (Climent-Soler *et al.* 2009). Still on the River Axe, rates of rise and their potential changes with land use have also been studied (Archer *et al.* 2010).

Nevertheless, the lack of an easily-accessible and quality-assured sub-daily flow dataset hampers the capability of large-scale national studies. In the U.S., such a dataset (Showstack 2007) has facilitated studies at national and regional scales, for instance: the flashiest catchment types and cities prone to flash floods have been identified in the continental U.S., with good matches found between places prone to floods and flood fatalities (Smith & Smith 2015), and a study of the seasonality of floods, allowing the correlation of location, timing, and drivers across the continental U.S. (Villarini 2016).

Furthermore, works on UK future hydrology have pointed to an overall increase of extreme events, with drought increases being more significant than floods (Collet *et al.* 2018). However, these studies use a daily timescale, while hourly rainfall extremes have been shown to increase at a higher rate than daily events (Guerreiro *et al.* 2018; Fowler *et al.* 2021); consequently, these studies might be underestimating flood events. A sub-daily national flow dataset could improve our understanding of future floods in the UK, showing the benefits of producing an open-source, quality-assured dataset.

Issues with the data – mismatches

Some mismatches have been manually checked, aiming to identify patterns in the dataset and to understand if there is an identifiable ‘truer’ value. There is a high variability in the reason for mismatches, some examples: failure of the instrument in high flows, such as a shaft encoder slip, rectified in one file but not in the other (NRFA station 33015–2003-01-04); manual modifications that were done in one time series and not in the other, such as vertical shifts (NRFA station 47008–2003-02-17), more often recorded in the 15 min time series and not in the irregular one, as in the sampled station; typos in dates (NRFA station 52014–1997-10-01); values that were computed but still unchecked in one sheet, while having been checked in the other (NRFA station 52006 –2000-November); and rating curve information mistranslations, in which the same station had non-mismatched level data and mismatched flow data (NRFA station 8426–2005-01-07). After 2003, not only are the number of mismatches and mistakes in the data greatly reduced, but also the errors are more systematic and detectable. For instance, NRFA station 47007–2004 to 2015 has one sheet with complete data, matching NRFA POT and AMAX flows while the second has missing data/not matching NRFA.

Additionally, the irregular time series often has repeated date-time values with different flows/levels. They reflect recordings that were taken at smaller uneven time-step intervals, with two recordings in the same minute. In most cases there is no indication that these records are incorrect, with a ‘good’ quality code by the agency and no visible issues when the flood wave is plotted. In the digitization process, for the purpose of depicting flood waves more accurately, charted data have sometimes been digitized in smaller than 15-min intervals. When the change was abrupt, the same time-step can have two different flow records. Nevertheless, in some cases, there will be an abrupt change in the magnitude of the flow, accompanied by a change in the quality code, from good to suspect, indicating errors in the measurements (NRFA station 68007–1992-09-18).

Regarding NRFA AMAX mismatches, even though both time series originate from the same 15-min dataset, the NRFA AMAX archive has additional quality-control checks and is regularly checked for the identification and removal of ‘flawed’ data. Before being part of the NRFA AMAX archive, the 15-min flow and level stations go through: (1) A selection process within the environmental agencies, aiming to remove stations that have unreliable values for high flows; (2) A verification of the quality of these stations periodically, e.g., SEPA verifies the level stations on a monthly basis and flow stations annually; (3) Additional quality-control by the NRFA, aiming for consistency in the AMAX and POT values; (4) Following the NRFA Quality Control (QC), some stations will be discarded while others will have their values edited to better reflect reality. Therefore, the mismatches between the NRFA AMAX values and the 15-min records are indicators that after further QC the peak flow of the station has been modified.

Integrating NRFA peak flow checks into the 15-min time series poses a challenge, since manual checks correct peak flow values but do not provide continuous flow event corrections. While extra quality-control efforts have been applied to these peaks, they cannot be systematically applied to the whole continuous record. The only automated outcome of these corrections is a flag in the 15-min record, indicating these discrepancies. Another challenge in the integration of the NRFA dataset to the continuous flow dataset available is the fluidity of the NRFA data, in which part of the stations are quality checked every year, and a new version, with corrected values, released to public.

Issues with the data – Other checks

Some stations in the 15-min dataset present values that are not physically realistic, that is, negative or higher than world record flow and level values. From initial manual inspection, these do not necessarily need to be discarded. Negative level values that are very close to 0 can reflect limitations in the instrument of measurement; while world record values could be a mistake in decimal places. Automatic flags to check and correct or discard these values are necessary.

A second quality step is using summary statistics, e.g., mean, min, max; and hydrologically-relevant indexes such as day of minimum/maximum streamflow (Gudmundsson *et al.* 2018), to identify potentially suspect gauges. Finally, an analysis of homogeneity using statistical tests (Gudmundsson *et al.* 2018; Crochemore *et al.* 2020) and the identification of high truncated values by checking streaks of repeated values (Lewis *et al.* 2021), can be used to further identify data-quality errors.

Continuous quality checks are performed manually on 15-min data in the UK by the NRFA and the agencies. Furthermore, projects such as the 2016 update to the National Risk Assessment of inland flooding risk (Aldridge *et al.* 2017) have produced occasional quality checks at a national level. Nevertheless, these have not produced a quality-control methodology that is nationally applicable, open-source, easily updatable and verifiable.

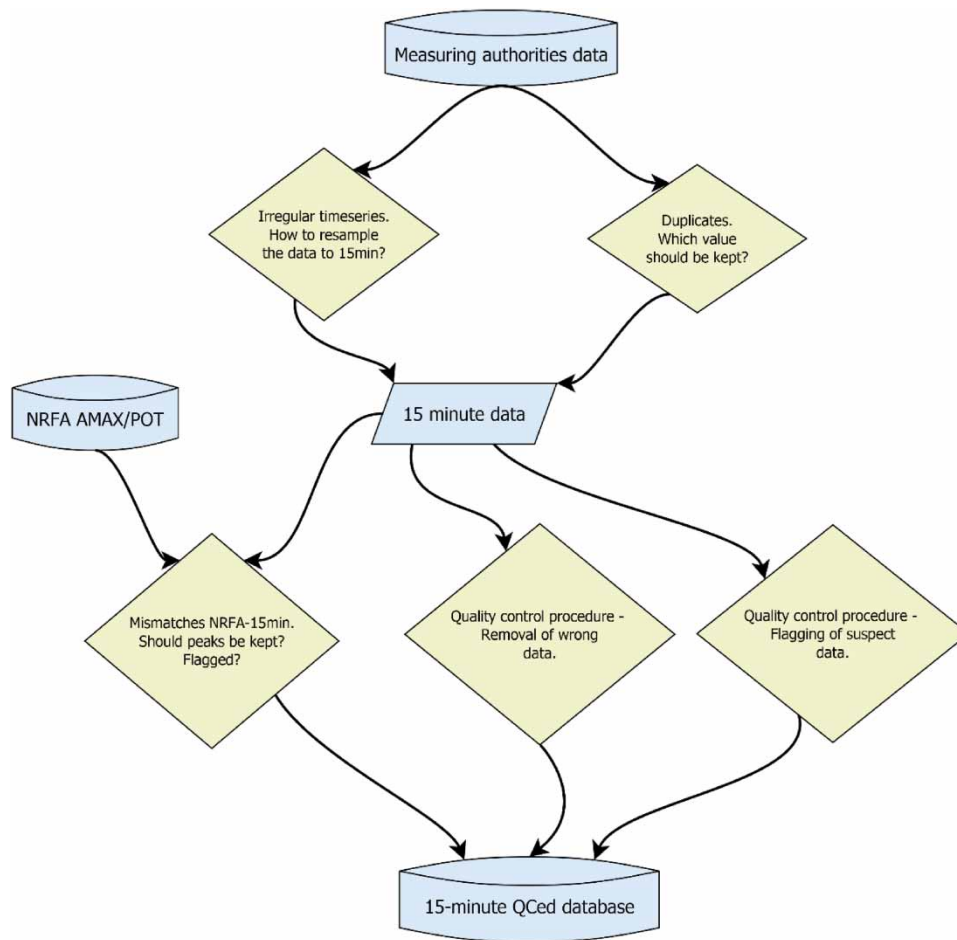


Figure 3 | Framework to quality control the datasets of the measuring authorities.

The 15-min dataset needs governance and curation

For a reliable 15-min QCed dataset, decisions on which values to keep and to flag as suspect will need to be taken (Figure 3). Furthermore, data governance and curation are imperative processes to assure data longevity and usability. Governance and curation processes should include: detailed documentation on every modification that has been done to the data; metadata with the maximum amount of information on the data source, e.g., provenance (EA, NRW, SEPA, and irregular or 15-min), station type, agency quality-control code; metadata information with the quality-control checks done on the data; the code used for processing the data; formatting of the data aiming for easy accessibility and understanding.

The benefits of having this information available include: the possibility of easily extending the dataset to more stations or time periods; identification of mistakes and incongruencies in the dataset on a station or systematic scale; possibility of modification of the dataset according to user needs such as a change in the resample timescale and addition of previously removed data; and a deeper user understanding of the capabilities and usability of the dataset.

CONCLUSIONS

The UK has sub-daily flow and level data recorded at a national level; nevertheless, no sub-daily national product is available to the public. Making such a dataset is fundamental for cutting-edge research, allowing the development of high-resolution continuous hydrological models. It is expected that the increase in the temporal resolution of models will help in understanding hydrological processes from the past and modelling potential changes from climate change, especially in high flow scenarios.

Making a comprehensible continuous sub-daily flow and level national dataset is a challenge. There is a lot of variation in how the data are kept by the different measuring authorities and according to the date of the records. To make the dataset trustworthy, the raw data must be cleaned and standardized; then, a national QC procedure should be applied to remove erroneous and to flag suspect data. Having an intelligible procedure is crucial to the upkeeping and improving the quality of the flow and level data. The measuring authorities time series are under constant review and modifications; hence, a procedure that allows data modification both ways is needed: wrong or suspect values detected by the automatic QC procedure should be translated to the agency time series; and in the other way, the detection of incorrect values by the authorities should also be translated to the QCed time series. From that perspective, data governance and curation with complete information on the procedures applied to the data can guarantee the dataset and the QC process longevity and reproducibility.

Finally, keeping the QC procedure open-source, verifiable and changeable is important in the optic of potential modifications to rating curves. A flow dataset, the most common input for hydrological model calibration, is not observationally based, being rather an estimation based on a rating curve. Rating curves are often changed based on observed changes in the channel. A flexible and updatable QC procedure could consider these future changes within the flow dataset.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

- Aldridge, T., Allan, R., Gouldby, B., Gunawan, O., Hunter, N., Lamb, R., Tawn, J. & Wood, E. 2017 *Spatial Joint Probability for FCRM and Strategic Assessments-Method Report*.
- Archer, D. & Fowler, H. 2021 *A historical flash flood chronology for Britain*. *Journal of Flood Risk Management* **14** (3), e12721.
- Archer, D. R., Climent-Soler, D. & Holman, I. P. 2010 *Changes in discharge rise and fall rates applied to impact assessment of catchment land use*. *Hydrology Research* **41** (1), 13–26.
- Arheimer, B. & Lindström, G. 2015 *Climate impact on floods: changes in high flows in Sweden in the past and the future (1911–2100)*. *Hydrology and Earth System Sciences* **19** (2), 771–784.
- Baker, M. 2016 *1,500 scientists lift the lid on reproducibility*. *Nature* **533** (7604), 452–454.
- Barbero, R., Fowler, H. J., Lenderink, G. & Blenkinsop, S. 2017 *Is the intensification of precipitation extremes with global warming better detected at hourly than daily resolutions?* *Geophysical Research Letters* **44** (2), 974–983.
- Bell, V. A., Kay, A. L., Jones, R. G., Moore, R. J. & Reynard, N. S. 2009 *Use of soil data in a grid-based hydrological model to estimate spatial variation in changing flood risk across the UK*. *Journal of Hydrology* **377** (3–4), 335–350.
- Beven, K. 1993 *Prophecy, reality and uncertainty in distributed hydrological modelling*. *Advances in Water Resources* **16** (1), 41–51.
- Beven, K. 2006 *A manifesto for the equifinality thesis*. *Journal of Hydrology* **320** (1–2), 18–36.
- Beven, K. 2019 *How to make advances in hydrological modelling*. *Hydrology Research* **50** (6), 1481–1494.
- Beven, K. 2021 *An epistemically uncertain walk through the rather fuzzy subject of observation and model uncertainties*. *Hydrological Processes* **35** (1), e14012.
- Beven, K. & Binley, A. 1992 *The future of distributed models: model calibration and uncertainty prediction*. *Hydrological Processes* **6**, 279–298.
- Beylich, M., Haberlandt, U. & Reinstorf, F. 2021 *Daily vs. hourly simulation for estimating future flood peaks in mesoscale catchments*. *Hydrology Research* **52** (4), 821–833.
- Blenkinsop, S., Lewis, E., Chan, S. C. & Fowler, H. J. 2017 *Quality-control of an hourly rainfall dataset and climatology of extremes for the UK*. *International Journal of Climatology* **37** (2), 722–740.
- Blenkinsop, S., Fowler, H. J., Barbero, R., Chan, S. C., Guerreiro, S. B., Kendon, E., Lenderink, G., Lewis, E., Li, X.-F., Westra, S., Alexander, L., Allan, R. P., Berg, P., Dunn, R. J. H., Ekström, M., Evans, J. P., Holland, G., Jones, R., Kjellström, E., Klein-Tank, A., Lettenmaier, D., Mishra, V., Prein, A. F., Sheffield, J. & Tye, M. R. 2018 *The INTENSE project: using observations and models to understand the past, present and future of sub-daily rainfall extremes*. *Advances in Science and Research* **15**, 117–126.
- Blöschl, G., Bierkens, M. F. P., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., Kirchner, J. W., McDonnell, J. J., Savenije, H. H. G., Sivapalan, M., Stumpp, C., Toth, E., Volpi, E., Carr, G., Lupton, C., Salinas, J., Széles, B., Viglione, A., Aksoy, H., Allen, S. T., Amin, A., Andréassian, V., Arheimer, B., Aryal, S. K., Baker, V., Bardsley, E., Barendrecht, M. H., Bartosova, A., Batelaan, O., Berghuijs, W. R., Beven, K., Blume, T., Bogaard, T., Borges de Amorim, P., Böttcher, M. E., Boulet, G., Breinl, K., Brilly, M., Brocca, L., Buytaert, W., Castellarin, A., Castelletti, A., Chen, X., Chen, Yangbo, Chen, Yuanfang, Chiffard, P., Claps, P., Clark, M. P., Collins, A.L., Croke, B., Dathe, A., David, P. C., de Barros, F. P. J., de Rooij, G., Di Baldassarre, G., Driscoll, J. M., Duethmann, D., Dwivedi, R., Eris, E., Farmer,

- W. H., Feiccabrino, J., Ferguson, G., Ferrari, E., Ferraris, S., Fersch, B., Finger, D., Foglia, L., Fowler, K., Gartsman, B., Gascoin, S., Gaume, E., Gelfan, A., Geris, J., Gharari, S., Gleeson, T., Glendell, M., Gonzalez Bevacqua, A., González-Dugo, M. P., Grimaldi, S., Gupta, A. B., Guse, B., Han, D., Hannah, D., Harpold, A., Haun, S., Heal, K., Helfricht, K., Herrnegger, M., Hipsey, M., Hlaváčiková, H., Hohmann, C., Holko, L., Hopkinson, C., Hrachowitz, M., Illangasekare, T.H., Inam, A., Innocente, C., Istanbuluoglu, E., Jarhani, B., Kalantari, Z., Kalvans, A., Khanal, S., Khatami, S., Kiesel, J., Kirkby, M., Knobon, W., Kochanek, K., Kohnová, S., Kolechkina, A., Krause, S., Kremer, D., Kreibich, H., Kunstmann, H., Lange, H., Liberato, M. L. R., Lindquist, E., Link, T., Liu, J., Loucks, D. P., Luce, C., Mahé, G., Makarieva, O., Malard, J., Mashtayeva, S., Maskey, S., Mas-Pla, J., Mavrova-Guirguinova, M., Mazzoleni, M., Mernild, S., Misstear, B. D., Montanari, A., Müller-Thomy, H., Nabizadeh, A., Nardi, F., Neale, C., Nesterova, N., Nurtaev, B., Odongo, V. O., Panda, S., Pande, S., Pang, Z., Papacharalampous, G., Perrin, C., Pfister, L., Pimentel, R., Polo, M. J., Post, D., Prieto Sierra, C., Ramos, M. H., Renner, M., Reynolds, J. E., Ridolfi, E., Rigon, R., Riva, M., Robertson, D. E., Rosso, R., Roy, T., Sá, J. H. M., Salvadori, G., Sandells, M., Schaeffli, B., Schumann, A., Scolobig, A., Seibert, J., Servat, E., Shafiei, M., Sharma, A., Sidibe, M., Sidle, R. C., Skaugen, T., Smith, H., Spiessl, S. M., Stein, L., Steinsland, I., Strasser, U., Su, B., Szolgay, J., Tarboton, D., Tauro, F., Thirel, G., Tian, F., Tong, R., Tussupova, K., Tyralis, H., Uijlenhoet, R., van Beek, R., van der Ent, R. J., van der Ploeg, M., Van Loon, A. F., van Meerveld, I., van Nooijen, R., van Oel, P. R., Vidal, J. P., von Freyberg, J., Vorogushyn, S., Wachniew, P., Wade, A. J., Ward, P., Westerberg, I. K., White, C., Wood, E. F., Woods, R., Xu, Z. & Yilmaz, K. K. & Zhang, Y. 2019 [Twenty-three unsolved problems in hydrology \(UPH\)–a community perspective](#). *Hydrological Sciences Journal* **64** (10), 1141–1158.
- Chen, B., Krajewski, W. F., Liu, F., Fang, W. & Xu, Z. 2017 [Estimating instantaneous peak flow from mean daily flow](#). *Hydrology Research* **48** (6), 1474–1488.
- Climent-Soler, D., Holman, I. P. & Archer, D. R. 2009 [Application of flow variability analysis to identify impacts of agricultural land-use change on the river axis, southwest England](#). *Hydrology Research* **40** (4), 380–393.
- Collet, L., Harrigan, S., Prudhomme, C., Formetta, G. & Beevers, L. 2018 [Future hot-spots for hydro-hazards in Great Britain: a probabilistic assessment](#). *Hydrology and Earth System Sciences* **22** (10), 5387–5401.
- Coxon, G., Freer, J., Westerberg, I. K., Wagener, T., Woods, R. & Smith, P. J. 2015 [A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations](#). *Water Resources Research* **51** (7), 5531–5546.
- Coxon, G., Addor, N., Bloomfield, J. P., Freer, J., Fry, M., Hannaford, J., Howden, N. J. K., Lane, R., Lewis, M., Robinson, E. L., Wagener, T. & Woods, R. 2020 [CAMELS-GB: hydrometeorological time series and landscape attributes for 671 catchments in Great Britain](#). *Earth System Science Data* **12** (4), 2459–2483.
- Crochemore, L., Isberg, K., Pimentel, R., Pineda, L., Hasan, A. & Arheimer, B. 2020 [Lessons learnt from checking the quality of openly accessible river flow data worldwide](#). *Hydrological Sciences Journal* **65** (5), 699–711.
- Environment Agency 2023 [Hydrology Data Explorer](#). Available from: <https://environment.data.gov.uk/hydrology/explore> (accessed 21 September 2023).
- Fosser, G., Murphy, J., Chan, S., Clark, R., Harris, G., Lock, A., Lowe, J., Martin, G., Pirret, J., Roberts, N., Sanderson, M., Tucker, S., Hoskins, B., Kjellström, E., Schär, C. & van den Hurk, B. 2019 [UKCP Convection-Permitting Model Projections: Science Report](#).
- Fowler, H. J., Lenderink, G., Prein, A. F., Westra, S., Allan, R. P., Ban, N., Barbero, R., Berg, P., Blenkinsop, S., Do, H. X., Guerreiro, S., Haerter, J. O., Kendon, E. J., Lewis, E., Schaer, C., Sharma, A., Villarini, G., Wasko, C. & Zhang, X. 2021 [Anthropogenic intensification of short-duration rainfall extremes](#). *Nature Reviews Earth and Environment* **2** (2), 107–122.
- Gudmundsson, L., Do, H. X., Leonard, M. & Westra, S. 2018 [The global streamflow indices and metadata archive \(GSIM\)-Part 2: quality control, time-series indices and homogeneity assessment](#). *Earth System Science Data* **10** (2), 787–804.
- Guerreiro, S. B., Fowler, H. J., Barbero, R., Westra, S., Lenderink, G., Blenkinsop, S., Lewis, E. & Li, X. F. 2018 [Detection of continental-scale intensification of hourly rainfall extremes](#). *Nature Climate Change* **8** (9), 803–807.
- Hannaford, J., Mackay, J., Ascot, M., Bell, V., Chitson, T., Cole, S., Counsell, C., Durant, M., Facer-Childs, K., Jackson, C., Kay, A., Lane, R., Mansour, M., Moore, R. J., Parry, S., Rudd, A., Simpson, M., Turner, S., Wallbank, J., Wells, S. & Wilcox, A. 2022 [Hydrological Projections for the UK, Based on UK Climate Projections 2018 \(UKCP18\) Data, From the Enhanced Future Flows and Groundwater \(eFLaG\) project](#).
- Haxton, T., Crooks, S., Jackson, C. R., Barkwith, A. K. A. P., Kelvin, J., Williamson, J., Mackay, J. D., Wang, L., Davies, H., Young, A. & Prudhomme, C. 2012 [Future Flows Hydrology Data](#).
- Hettiarachchi, S., Wasko, C. & Sharma, A. 2019 [Can antecedent moisture conditions modulate the increase in flood risk due to climate change in urban catchments?](#) *Journal of Hydrology* **571** (January), 11–20.
- Kay, A. L. 2021 [Simulation of river flow in Britain under climate change: Baseline performance and future seasonal changes](#). *Hydrological Processes* **35** (4), 1–10.
- Kjeldsen, T. R. 2007 [The Revitalised FSR/FEH Rainfall-Runoff Method](#).
- Lane, R. A., Coxon, G., Freer, J. E., Wagener, T., Johnes, P. J., Bloomfield, J. P., Greene, S., Macleod, C. J. A. & Reaney, S. M. 2019 [Benchmarking the predictive capability of hydrological models for river flow and flood peak predictions across over 1000 catchments in Great Britain](#). *Hydrology and Earth System Sciences* **23** (10), 4011–4032.
- Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G. & Dadson, S. J. 2021 [Benchmarking data-driven rainfall-runoff models in Great Britain: a comparison of long short-term memory \(LSTM\)-based models with four lumped conceptual models](#). *Hydrology and Earth System Sciences* **25** (10), 5517–5534.
- Lewis, E., Birkinshaw, S., Kilsby, C. & Fowler, H. J. 2018a [Development of a system for automated setup of a physically-based, spatially-distributed hydrological model for catchments in Great Britain](#). *Environmental Modelling and Software* **108**, 102–110.

- Lewis, E., Quinn, N., Blenkinsop, S., Fowler, H. J., Freer, J., Tanguy, M., Hitt, O., Coxon, G., Bates, P. & Woods, R. 2018b A rule based quality control method for hourly rainfall data and a 1 km resolution gridded hourly rainfall dataset for Great Britain: CEH-GEAR1hr. *Journal of Hydrology* **564**, 930–943.
- Lewis, E., Pritchard, D., Villalobos-Herrera, R., Blenkinsop, S., McClean, F., Guerreiro, S., Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Rustemeier, E. & Fowler, H. J. 2021 Quality control of a global hourly rainfall dataset. *Environmental Modelling and Software* **144**, 105169.
- Massari, C., Brocca, L., Moramarco, T., Tramblay, Y. & Didon Lescot, J. F. 2014 Potential of soil moisture observations in flood modelling: estimating initial conditions and correcting rainfall. *Advances in Water Resources* **74**, 44–53.
- McMillan, H., Krueger, T. & Freer, J. 2012 Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality. *Hydrological Processes* **26** (26), 4078–4111.
- McMillan, H. K., Westerberg, I. K. & Krueger, T. 2018 Hydrological data uncertainty and its implications. *Wiley Interdisciplinary Reviews: Water* **5** (6), e1319.
- McMillan, H. K., Coxon, G., Sikorska-Senoner, A. E. & Westerberg, I. K. 2022 Impacts of observational uncertainty on analysis and modelling of hydrological processes: preface. *Hydrological Processes* **36** (2), e14481.
- NRW 2016 *Natural Resources Wales' API Portal*.
- Prein, A. F., Rasmussen, R. M., Ikeda, K., Liu, C., Clark, M. P. & Holland, G. J. 2017 The future intensification of hourly precipitation extremes. *Nature Climate Change* **7** (1), 48–52.
- Prosdocimi, I., Kjeldsen, T. R. & Miller, J. D. 2015 Detection and attribution of urbanization effect on flood extremes using nonstationary flood-frequency models. *Water Resources Research* **51** (6), 4244–4262.
- Robson, A. J. & Reed, D. W. 1999 *Statistical Procedures for Flood Frequency Estimation. Volume 3 of the Flood Estimation Handbook*. Centre for Ecology & Hydrology, Wallingford, UK.
- SEPA 2022 *SEPA Time Series Data Service (API)*.
- Sharma, A., Hettiarachchi, S. & Wasko, C. 2021 Estimating design hydrologic extremes in a warming climate: alternatives, uncertainties and the way forward. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **379** (2195), 20190623.
- Showstack, R. 2007 Online database for instantaneous streamflow data. *Eos, Transactions American Geophysical Union* **88** (48), 523–523.
- Smith, B. K. & Smith, J. A. 2015 The flashiest watersheds in the contiguous United States. *Journal of Hydrometeorology* **16** (6), 2365–2381.
- Stagge, J. H., Rosenberg, D. E., Abdallah, A. M., Akbar, H., Attallah, N. A. & James, R. 2019 Assessing data availability and research reproducibility in hydrology and water resources. *Scientific Data* **6**, 190030.
- Tanguy, M., Dixon, H., Prosdocimi, I., Morris, D. G. & Keller, V. D. J. 2021 *Gridded Estimates of Daily and Monthly Areal Rainfall for the United Kingdom (1890–2019) [CEH-GEAR]*.
- Villarini, G. 2016 On the seasonality of flooding across the continental United States. *Advances in Water Resources* **87**, 80–91.
- Wagener, T., Dadson, S. J., Hannah, D. M., Coxon, G., Beven, K., Bloomfield, J. P., Buytaert, W., Cloke, H., Bates, P., Holden, J., Parry, L., Lamb, R., Chappell, N. A., Fry, M. & Old, G. 2021 'Knowledge gaps in our perceptual model of Great Britain's hydrology'. *Hydrological Processes* **35** (7), e14288.
- Wallingford HydroSolutions 2016 *WINFAP 4 QMED Linking Equation*.
- Wasko, C. & Nathan, R. 2019 Influence of changes in rainfall and soil moisture on trends in flooding. *Journal of Hydrology* **575** (May), 432–441.
- Wasko, C. & Sharma, A. 2017 Global assessment of flood and storm extremes with increased temperatures. *Scientific Reports* **7** (1), 1–8.

First received 30 March 2023; accepted in revised form 29 September 2023. Available online 13 October 2023