Environmental Modelling and Software 124 (2020) 104598

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: http://www.elsevier.com/locate/envsoft

SRS-GDA: A spatial random sampling toolbox for grid-based hydro-climatic data analysis in environmental change studies

Han Wang, Yunqing Xuan *

Zienkiewicz Centre for Computational Engineering, College of Engineering, Swansea University Bay Campus, Fabian Way, Swansea, SA1 8EN, UK

ARTICLE INFO

Keywords:
Spatial random sampling
Grid-based data analysis
Environment change
MATLAB toolbox
Open source software

ABSTRACT

We present in this paper the development of a new, open-source MATLAB toolbox SRS-GDA that aims to provide random spatial sampling of grid-based hydro-climatic datasets for environmental change studies. This toolbox addresses the needs of quantifying how hydro-climatic responses, which are often driven by grid-based forcing datasets such as climate model projections, vary with location and scale. The toolbox can be used to carry out random spatial sampling of grid-based quantities with various constraints: shape, size, location, dominant orientation and resolution. A case study of a large dataset, the GEAR rainfall dataset is supplied to demonstrate the typical uses case of this toolbox. The provision of the toolbox for downloading together with the sample data is also presented.

1. Introduction

Research in environmental changes has been increasingly relying on computer models driven by external forcing field and conditions that can represent changing factors such as temperature, precipitation and land uses (Erler et al., 2019; Alamou et al., 2017). Historically, the inputs used to drive these models were often relatively scarce, and only available at limited number of locations. Data collection was often restricted by technical conditions, instruments and means of storage. To make full use of such finite data, many methods have been proposed and applied. In terms of rainfall data, there are many methods for translating point rainfall records which are usually collected from hydrological gauges or stations to the basin rainfall. For example, the Areal Reduction Factor (ARF) has been widely used, possibly under different names in different countries (Weather Bureau, 1958; NERC, 1977). More recently, however, with the rapid advances in environmental monitoring technology, spatially disaggregated, grid-based hydro-climatic datasets have become gradually available with steady improvements in both accuracy and spatial-temporal resolutions. A typical case is the NIMROD weather radar system deployed by the UK Met Office which can provide up to 1km/5min precipitation distribution over the country (Golding, 1998; Fairman et al., 2017). Similarly, satellite-borne observations, such as the Global Precipitation Measurements (GPM; Islam et al., 2014; Ning et al., 2016) can now provide large scale coverage of the precipitation coverage in near real-time. Many environmental models nowadays are also tuned to make use of those new, grid-based, high resolution datasets, e.g. the Grid-to-Grid version of the PDM model has been developed and operationally used by the Environment Agency in the UK (Cole and Moore, 2008; Kay et al., 2009).

Another important source of external forcing data is model simulated hydro-climatic fields. In this case, rainfall, temperature as well as soil moisture fields generated by numerical weather models or climate models can be used to drive other model simulations. Practices of using the so-called coupled model approach started to gain momentum in the early 2000's when numerical weather models and climate models were able to produce simulation with high enough spatial resolutions, e.g. at tens of kilometres. As such, there have been plenty of studies since then, such as Bauer et al. (2015), Moufouma-Okia & Jones (2015) and many more inspired by the Hydrological Ensemble Prediction Experiments (HEPEX; Schaake et al., 2007) initiative. Datasets such as the ERA40 (Uppala et al., 2005) have been widely used since then. Although these datasets are not originally produced over sets of grids, or at least not the commonly recognised types of grids; they often are interpolated onto regular grids nevertheless in order to facilitate further analysis and to be used as other model inputs. For instance, global numerical weather models tend to use the Gaussian Grids, e.g., EAR40 grids. Local area model (LAM), such as the Weather Research and Forecasting model (WRF; Skamarock et al., 2001) uses regular grids spatially but does so only on a projected plane.

The importance and popularity of using those grid-based forcing data

* Corresponding author.
E-mail address: y.xuan@swansea.ac.uk (Y. Xuan).

https://doi.org/10.1016/j.envsoft.2019.104598
Received 19 July 2019; Received in revised form 23 October 2019; Accepted 3 December 2019
Available online 5 December 2019

are underlined by the needs of many climate change impact studies where climate projections, such as those from the Coupled Model Intercomparison Project (Giorgetta et al., 2013; Covey et al., 2003), are normally provided over a set of regular longitude/latitude grids over the globe. To better facilitate the community in using these grid based data and encourage the interoperability among models, the Network Common Data Format (NETCDF, Rew and Davis, 1990) has become the de-facto standard in climate change impact studies, although other traditional formats such as GRIdded Binary (GRIB, Rutledge et al., 2006) or Hierarchical Data Format (HDF, Duane et al., 2000) are well supported as well.

In the context of using grid-based hydro-climatic datasets for providing external forcing field, an important step is to understand, quantify and if possible, to correct the errors and/or bias in these fields. The spatially variant nature of these data remains as the centre of the process. For example, Rojas et al. (2011) applied a statistical bias correction to improve the regional climate model (RCM)-driven climate simulations across the Europe; Rabiei and Haberlandt (2015) proposed to merge rain gauge measurements and weather radar data which is grid-based data by bias correction. Specifically for weather radar adjustment, many algorithms such as the Mean Field Bias (MFB) method and the Kriging with External Drift (KED) method, adjust the radar data solely by a multiplicative factor which does not vary spatially; however, more recently the Conditional Merge algorithm introduced by Sinclair and Pegram (2005) and implemented by Guenzi et al. (2016), considers the spatial impacts by conditioning the gauge adjustment on the radar precipitation values at gauge locations (Silver et al., 2019).

Apart from being used as inputs to the models, the grid-based hydro-climatic datasets are also a foundation to support further analyses on environment change both spatially and temporally. It is not surprising that nearly all published studies in this field have been done so on grid-based datasets, e.g., Du and Zhang (2019) identified the spatiotemporal variations and trends of precipitation and streamflow extremes in the Xiang river basin with gridded data of resolution 0.5° and concluded that intensified summer extreme precipitation occurs mainly in the upper and middle of the basin and extreme streamflow has an increasing trend at the same region; Fairman et al. (2017) analysed the climatology of size, shape and intensity of grid-based precipitation features over Great Britain and Ireland; more application on grid-based data can be seen in Drusch et al. (2004); Thorndahl et al. (2017); Chen et al. (2015).

It is clear from the above examples that the grid-based hydro-climatic data have spatial patterns and characteristics with regards to certain changing factors that need to be diagnosed. Such diagnosis, without exception, is done over analysing targeted variable(s) and/or their combinations sampled spatially within predefined boundaries such as political regions (Bell et al., 2009) and river catchments (Monteiro et al., 2016). Further, to understand the random nature of the errors and uncertainties associated with the spatial data, the Monte-Carlo simulation approach is commonly used together with geo-statistical stochastic simulation for uncertainty quantification. A simple procedure is to perform simulations of points (can be data or events) randomly distributed in the predefined area, calculate the empirical distribution function of such inter-point distances in each case and then obtain further values of the statistic by goodness of fit (Besag and Diggle, 1977). Following this approach, some applications have been published, e.g., Smith and Cheeseman (1986); Xu et al. (2005) and Wu et al. (2018); however, application on hydro-climatic grid-based data remains scarce and many previous studies on spatial distributions of hydro-climatic variables were conducted over predefined areas.

Apparently, the substantial overhead of computer programming of spatial random sampling over often large-size hydro-climatic datasets has affected researchers' capacity of studying spatial-temporal variation of climatic features. To address this issue, we developed a Spatial Random Sampling toolbox for Grid-based Data Analysis (SRS-GDA) which can generate arbitrary samples from any grid-based dataset automatically. As an Open-source MATLAB toolbox, it can assist users in spatial random sampling with various constraints such as shape, size, location, dominant orientation and resolution. In the field of environmental change impact studies where the spatial properties of grid-based datasets remain as the focus, this toolbox addresses the needs of quantifying how hydro-climatic responses vary with location and scale. The grid size used by the SRS-GDA toolbox can be defined in line with any resolution of the base grid map. To increase the applicability of this toolbox, users can customise various sampling conditions and their combinations which can be directly applied to many environmental change studies.

This paper is structured as follows: first, a brief introduction of the study background and the main objective are provided, followed by the presentation of the methodology section. An example using case of analysing hydro-climatic extremes, i.e. precipitation over Great Britain using the GEAR dataset is provided to demonstrate the application of the toolbox. Finally, discussion on further applicability and availability of the model is presented.

2. The design and implementation of the SRS-GDA toolbox

The main aim of the SRS-GDA toolbox is to enable random spatial sampling of grid-based data within a pre-defined Region of Interest (ROI) of different sizes, shapes, locations and resolutions. The sampling procedure starts with a user-supplied grid dataset with spatial reference. It is also common to have an overall boundary (OB) from which the sampling is to be conducted, as many grid datasets have coverage normally much larger than that of the user's interest, such as the General Circulation Model (GCM) output around the globe. Normally, the OB should be set large enough for studying how the variation of locations can affect certain quantities represented by an ROI.

The randomisation of the sampling process is manifested by the ways of how the ROI is constructed:

1) Randomisation of the shape of the ROI. The shape of an area often plays an important role in various applications. For example, in hydrology, a so-called donor catchment is often desired to have a shape analogous to that of the ungauged, target one. Understandably, this process sets to be the most complex one in the SRS-GDA toolbox. We offer two options with regards to whether the shape of the ROI is of concern: the shape-unconstrained sampling which randomises the shape of ROI; and the shape-constrained sampling that makes use of a predefined geometric shape supplied by the user e.g. a polygon at a given scale. A special case is point or single-grid sampling whose ROI reduces to a single grid. This is also useful, for example, when studying the variation of point-measured quantities.
2) Randomisation of the location of the ROI. The location of an ROI can be varied by changing the coordinates of its centroid (for predefined ROI's) or its origin (for randomly generated ROI's). This operation is done by randomly setting a point or grid within the OB as the new location for the ROI to be moved to. An extra step is usually applied to ensure the entire region of the ROI falling within the OB.
3) Randomisation of the size of the ROI. Variation of the ROI size can help users to identify whether the aggregated data value over an area exhibits notable behaviour. A typical case, for example, is to study the extreme value distribution of a hydrometeorological variable – temperature or precipitation, at regional, national and global scale. This operation depends on whether the ROI is shape-constrained or not. If a predefined shape is used, a 'buffering' operation (Chang, 2008) is used to either increase or reduce the size whilst maintaining the shape unchanged; whereas for a shape-unconstrained case, the desired ROI is randomly produced with a given location and specified size.

These three operations can be combined to achieve the various levels of randomisation required by users. The implementation of the toolbox involves a series of steps that are described below and shown in Fig. 1

which includes: (a) Grid map generation which sets the overall boundary (OB) spatial coverage constraint and the resolution for the study (sampling) area; (b) Sampling setup that determines whether one or more constraints are used and sets the corresponding values and/or features, for example, location (fixed or floated), shape-unconstrained or shape-constrained, size fixed or not etc. and (c) Sampling processing and validation which are automatically carried out by the SRS-GDA toolbox based on the OB grid map and the constraint setups with extra filters applied to the results depending on extra conditions where appropriate.

2.1. Generating the grid-based overall boundary (OB) map

As mentioned previously, the underlying dataset normally comes with a coverage larger than that of users' interest. In other words, a subset based on an OB needs to be produced. This OB needs to be specified by the user, e.g., by using either a raster file or a vector based map such as shapefiles that define the boundary. If no OB is specified, the entire coverage of the underlying grid dataset will be used to conduct the sampling process. It should be mentioned that the sampling process often happens inside the OB. However, different from OB, the boundaries specified by the ROI's are deemed to be restrictive and arbitrary as far as a natural process is concerned, such as rainfall and wind speed.

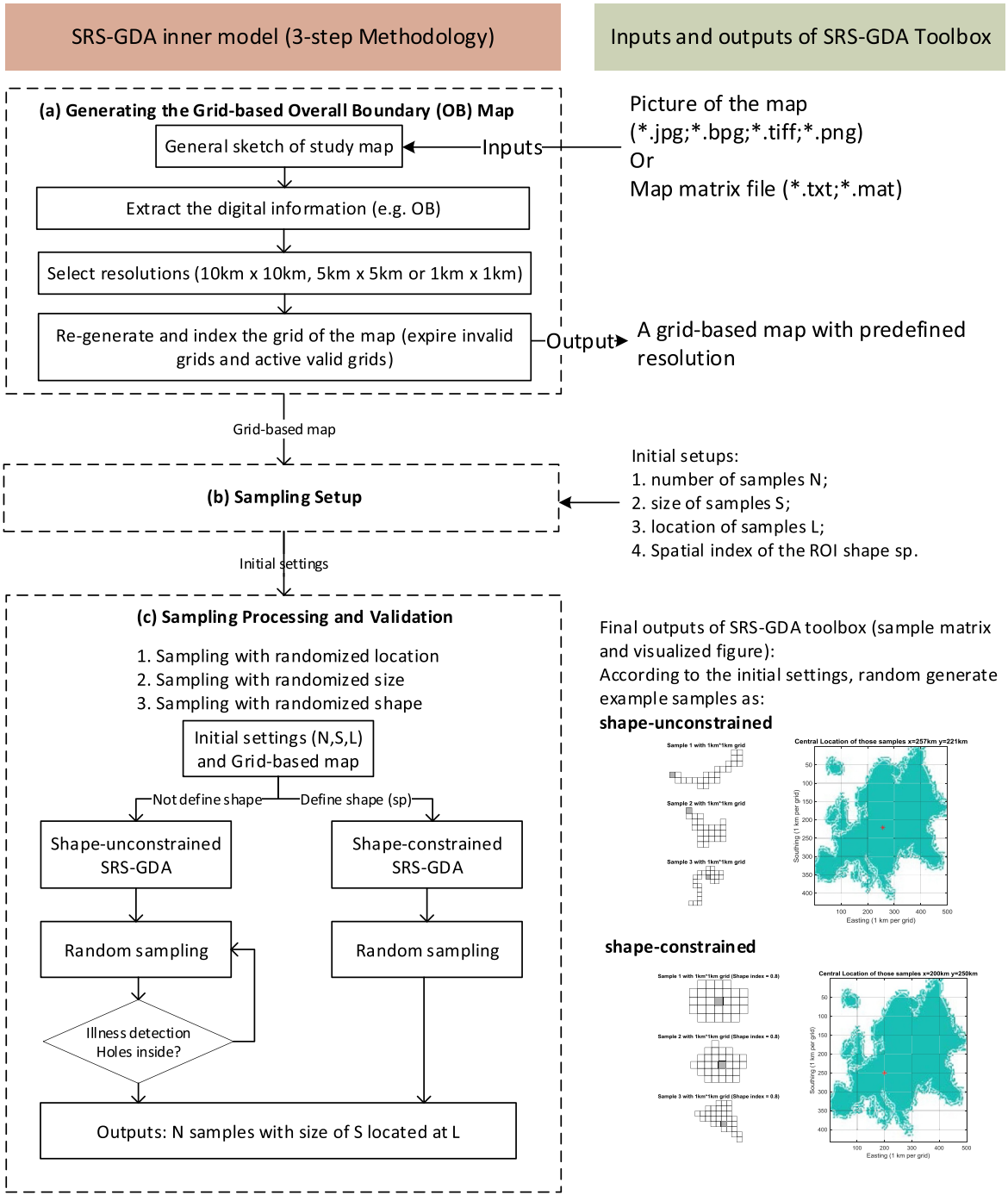

Fig. 1. The three basic steps and the corresponding inputs and outputs of the SRS-GDA toolbox.

The logic behind sampling ROI in OB is because many times only the quantity of certain hydro-climatic variables falling in such given boundary is of concern, for example, rainfall over the urban area of a city is a key element for urban drainage design.

A grid-based map is then generated by rasterising the OB (if it comes as a vector map) using the same projection and grid resolution as the underlying dataset. The grids inside the OB are regarded as valid grids while those outside are invalid grids. Once this is completed, the toolbox will automatically exclude those invalid grids and activate the valid grids. For example, in the example case given in this paper, the National Grid Reference (NGR, Ordnance Survey, 1946) is used to refer to the coordinates of the grids of the GEAR dataset. The base map is processed to distinguish ocean (so called invalid grids outside the GB boundary) and land (so called valid grids inside the GB boundary). It is also further refined to have several versions with different spatial resolutions which are normally multiples (exact divisions) of the grid size of the underlying dataset. These refined OB's will be used for further study on aggregation (upscaling) and disaggregation (downscaling). The toolbox provides three resolutions to match the underlying grid dataset: 1 km × 1 km, 5 km × 5 km and 10 km × 10 km for user applying. And the base maps of the GB are produced with these three resolutions respectively, as shown in Fig. 5 where 1 km × 1 km is chosen for demonstrating the example case for being consistent with the resolution of dataset (details in 3.1).

In addition to setting the OB, another important task at this step is to spatially index the data grids and label those that contain valid data. From now on, all subsequent spatial sampling is conducted over (or within, to be more precise) the base map.

2.2. Sampling setup

There are four initial settings (also seen in Fig. 1b) that need to be specified before starting the sampling process which are:

1) Total number of samples required;
2) The desired location of the samples, which is only applicable in the case where users wish to fix the location while randomising other properties such as shapes and sizes;
3) Sample size in the unit of km^2 which is translated into numbers of grids at the finest grid resolution used; Note that this is only required if a size-constrained sampling is desired;
4) Spatial index of the ROI shape (i.e., samples) which is needed when a shape-constrained sampling is required. In this case, the ROI shapes are randomly generated as convex hulls having their spatial index (*sp*) value set by the user. In the case of shape-unconstrained sampling, the shape of the ROI's will be randomised. The spatial index (*sp*) is defined to indicate dominant spatial extension direction, e.g., north-south or west-east:

$$sp = \frac{D_{NS}}{D_{WE}} \tag{1}$$

where D_{NS} and D_{WE} refer to the north-south dimension and the west-east dimension of a sample (represented by a matrix). The reason of having *sp* as an attached indictor is that in many climate studies, the direction of an area (such as a river catchment) plays a crucial role in determining the amount of quantity, such as rainfall (Viviroli et al., 2003; Svensson and Rakhecha, 1998). Obviously, other indexes, such as the direction of the major axis, can be easily defined if required.

2.3. Sampling processing and validation

This is the final step (Fig. 1c) where samples are generated according to the initial settings. The methods discussed below correspond to the three main functions of SRS-GDA toolbox.

- Sampling with randomised locations

This function randomly selects different locations to set the centroids of the samples within the OB base map. The sampling is relatively straightforward: first *x*- and *y*-coordinates are sampled from the range of the OB maps in the two directions using a joint uniform distribution $U(X, Y)$; followed by filtering out those samples that are not entirely within the OB.

- Sampling with randomised sizes

The second function is to randomly generate samples with different sizes, which is mainly used in the cases where the behaviour of aggregated quantity over the area of a sample is desired. Since the grid resolution A_{grid} (in km^2) is known, the size of sample A_{sample} can be translated into the number of grids $N_{grids\ of\ sample}$ of the ROI. The equation below shows the translation:

$$N_{grids\ of\ sample} = A_{sample} / A_{grid} \tag{2}$$

The variation of the area of the ROI (the sample) is realized by applying a 'buffering' operation while keeping the centroid location unchanged, i.e., it only increases or decreases the main axis of the sample proportionately. Fig. 2 shows an example of shape generation.

- Sampling with randomised shape of ROI: unconstrained and constrained

The third main function is to randomly generate samples in different shapes varying in both sizes and locations. Depending on the user's initial settings, this function can conduct both shape-unconstrained and shape-constrained sampling. In the former case, the location and the size of the sample (ROI) are both obtained from the two previous functions; for each combination of the location and the size, the shape is randomised using the size as a constraint. Two principles are applied in this process:

1) all grids should be interconnected, i.e. no isolated grids are allowed;
2) any growth must not go over the boundary set by the OB map.

The sampling starts at the given location and follows a random run to the neighbouring grid and records it until the number of grids equals $N_{grids\ of\ sample}$. All the grids covered by the path are selected to comprise the sample. An extra validation step is applied to remove samples with holes inside (the so-called ill-set samples) and rerun the process until the required number of samples is met, as presented in Fig. 3.

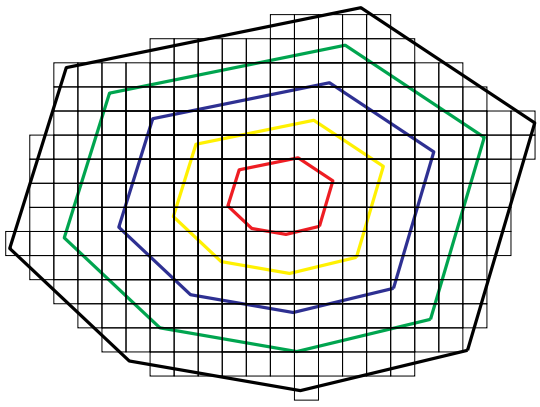

Fig. 2. The 'buffering' operation used to vary the ROI into difference sizes (shown here in different border colours). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

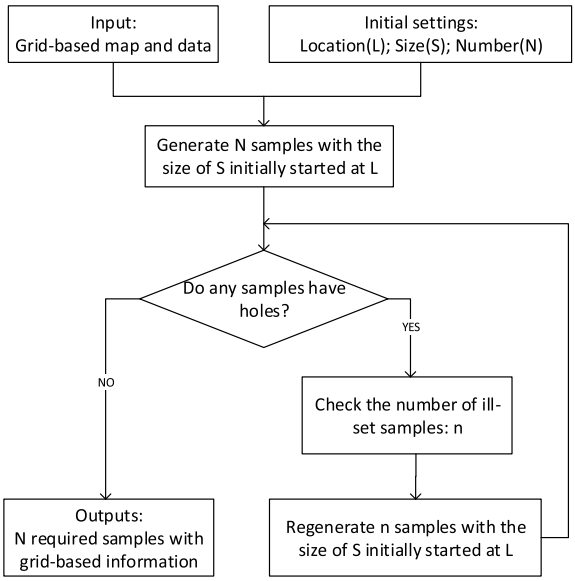

Fig. 3. The process of shape-unconstrained random sampling method with ill-sample detection and removal.

For the case of shape-constrained random sampling, it focuses on sampling with the shapes of convex polygons as seen in many hydrological catchments in environmental or climatic research. The working flow is shown in Fig. 4.

Unlike the shape-unconstrained method, the shape-constrained random sampling method produces more regular samples such as convex polygons. The main parameters such as the initial/centred location (L), sample size (S) and number (N) are the same as those required by the shape-unconstrained method. In addition, the shape-constrained method uses one more major parameter the spatial index (*sp*) as a further constraint. If required, three optional parameters can also be set to further refine the control of the polygon generation, i.e. the number of angles (usually is greater than or equal to 3); the irregularity that indicates how much variance there is in the angular distance of vertices with a range of 0–1; the spikiness which indicates how much variance there is in each vertex from the average radius with a range of 0–1. However, as in the setup of the main parameters, L, S, N and *sp*, specification of these additional parameters are not compulsory. Unless otherwise specified explicitly by the user, the toolbox automatically generates default values for them (e.g. irregularity = 0.3 and spikiness = 0.1) to avoid producing extremely weird (irregularity = 1) or sharp (spikiness = 1) polygons. Compared with the shape-unconstrained random sampling method, it runs substantially faster because there is no need for random walking to grow the grids nor does it have any possibility of producing ill-set areas.

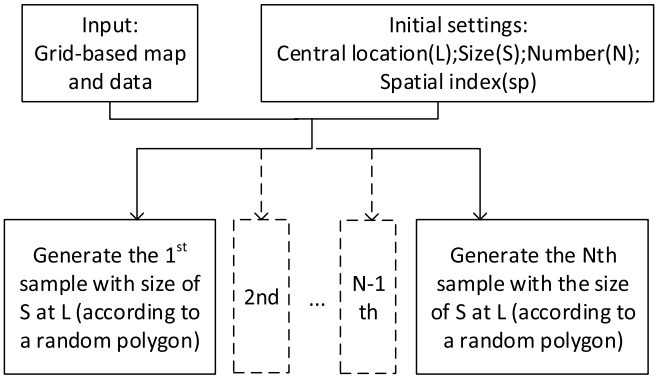

Fig. 4. Flowchart of shape-constrained sampling method.

3. An example application of the toolbox

3.1. Dataset

One of the motivations of this example is to investigate how areal rainfall extremes in terms of their distributions can vary with locations, size and shapes of the ROI. In fact, there has been consensus about the impact of the size of catchment when producing areal rainfall at certain return levels. This is normally acknowledged by applying a so-called Areal Reduction Factor (ARF, Bell, 1976) to the value obtained at the location of the centroid of the catchment. Whilst variation of hydro-climatic variables is commonly recognised to be associated with the climatology, impact of the locations as well as the shape of the catchment have not been fully studied in a quantitative way. In our case, the 1-km gridded estimates of daily rainfall for Great Britain are analysed using a map of Great Britain roughly sized as 700×1250 km^2. The rainfall estimates are derived from the Met Office national database of observed precipitation by using the UK rain gauge network. The natural neighbour interpolation methodology, including a normalization step based on average annual rainfall, was used to generate the daily estimates from 9am until 9am on the following day (Tanguy et al., 2016).

3.2. Application of the SRS-GDA toolbox

To be consistent with the precision of dataset, the OB base map is produced as the same grid size of 1 km^2. The production of the OB map undergoes two steps: first, a rough sketch of the boundary of Great Britain (GB) is used to generate grids with very coarse resolution set as 100 km^2. This is to ensure the boundary is properly covered. Secondly, the grid map is then refined by subdividing every grid with a number of smaller ones so that the grid resolution gradually increases to 5 km $\times$ 5 km and 10 km $\times$ 10 km, which allows for the detection and removal of those grids falling outside of the boundary. This process is shown in Fig. 5 including: (a) 638607 valid grids (marked as green) with the size of 1 km^2; (b) 9464 valid grids with the size of 25 km^2; (c) 2368 valid grids with the size of 100 km^2.

Meanwhile, the location of the sample in this example study is chosen to be in London with the coordinate of L = (520 km, 1070 km). Two random sampling methods, e.g. shape-unconstrained and shape-constrained, are used to generate 5 different samples (N = 5) at this location with the same size of 25 km^2. According to Eq. (2), the number of grids in each sample (S) is calculated as 25 km^2/1 km^2 = 25. N, L and S are the basic inputs for SRS-GDA toolbox.

3.2.1. Shape-unconstrained random sampling method

Table 1 presents the 5 different samples around the initial location **L** (grey grid) generated by the shape-unconstrained random sampling method. It can be observed that all samples have grids interconnected with no hole inside. However, the shapes of the sample can be very irregular as there is no requirement that they need to be a convex polygon which is used in the shape-constrained sampling method.

The shape-unconstrained sampling offers maximum freedom; however, it can inevitably introduce shapes with holes inside, which have to be rejected. Fig. 6 shows the steps involved to detect and remove those ill-set sample shapes: First the original sample is presented to the validation function (Fig. 6a) before it is converted into a binary image (Fig. 6b). Secondly, the inner area of the binary image is flooded to remove the potential holes which results in a hole-free image as shown in Fig. 6c. Finally, by comparing the areas of the two images, the location and the size of the hole(s) can be detected, which in turn triggers the removal process to discard the ill-set sample. In our test, the whole process of shape-unconstrained random sampling method takes 7.0 s on a low-configuration laptop to randomly generate five accepted samples

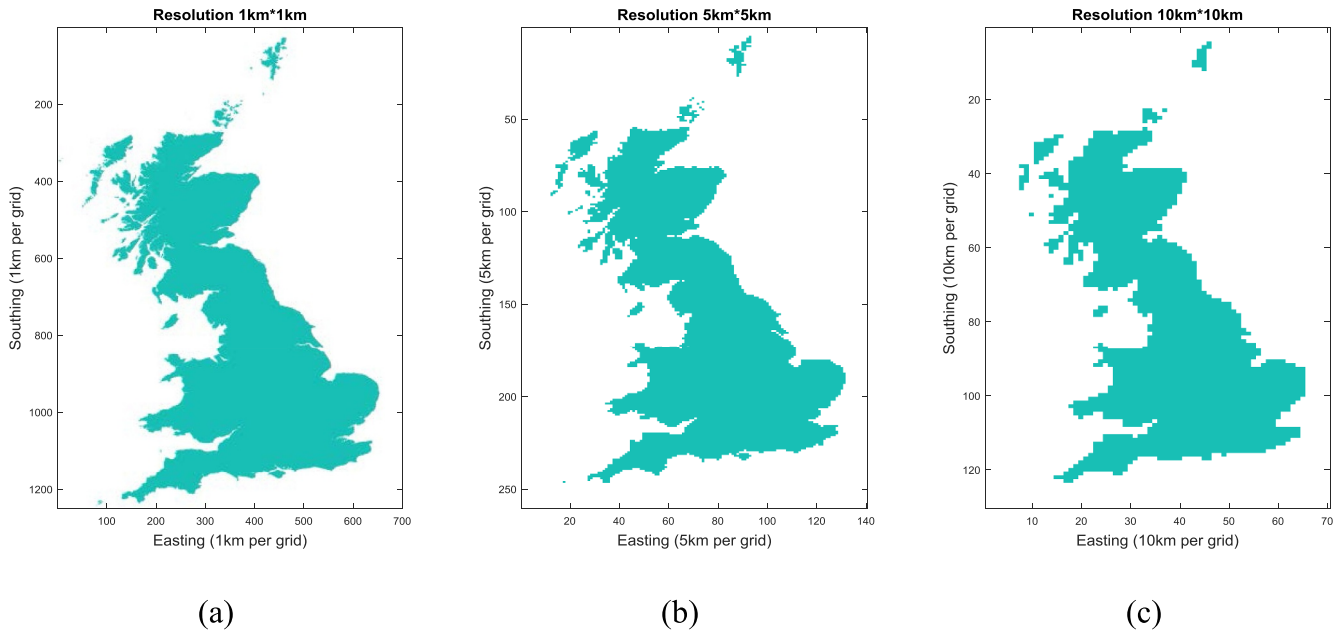

Fig. 5. General map of Great Britain with three resolutions: (a) 1 km*1 km (b) 5 km*5 km (c) 10 km*10 km. The difference in details and resolutions can be appreciated in the representation of the coast lines.

Table 1
Five example samples generated by shape-unconstrained sampling method.

	No.1	No.2	No.3	No.4	No.5
Samples					

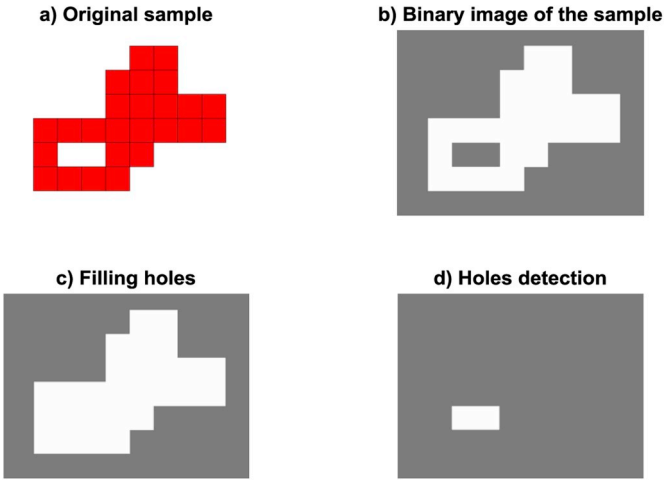

Fig. 6. The process of hole detection.

with sizes of 25 km^2 (specified as an initial constraint) while three samples are abandoned.

3.2.2. *Shape-constrained random sampling method*

Five samples at same location L (grey grid) generated by using shape-constrained random sampling method are shown in Table 2 with various spatial indexes *sp* defined by the toolbox. Comparing with those samples listed in Table 1, clearly the shapes are more regular here as convex polygons, which can be directly used to simulate hydrological catchments. The whole process is recorded to have finished in 2.0s on our test PC, which is shorter than that from the former method. However, the tests show that the larger size and number are, the more efficient and time-saving the shape-constrained method is, compared with the shape-unconstrained method in Table 3.

Fig. 7 summarises the steps taken for shape-constrained sampling starting with an arbitrary but convex polygon (with *sp*, irregularity and spikiness all set by the toolbox) set at the same location index L (grey grid).

The effect of the spatial index *sp* in the process of shape-constrained sampling is shown in Fig. 8 with larger values of *sp* having more north-

Table 2
Five example samples generated by shape-constrained sampling method.

	No.1	No.2	No.3	No.4	No.5
Samples					

Table 3
Comparison of the indicative speed of the two sampling methods: Method 1 the shape-unconstrained method and Method 2 the shape-constrained method. Note that the numbers are obtained on our test PC and for comparing the relative speed difference.

Number of Grids	Sampling Method	Number of Samples						
		5	10	20	45	60	100	150
25	Method 1	7.4s	13.8s	39.1s	2.3min	3.6min	9.3min	20.9min
	Method 2	2.2s	2.8s	3.0s	4.6s	5.6s	7.0s	10.0s
50	Method 1	18.1s	33.1s	1.8min	8.1min	29.4min	39.6min	1.4 h
	Method 2	2.1s	3.8s	4.7s	7.7s	10.7s	12.0s	20.6s
100	Method 1	50.8s	3.5min	28.1min	1.2 h	2.4 h	9.6 h	12.9 h
	Method 2	1.7s	3.1s	7.3s	11.4s	13.5s	23.0s	30.5s

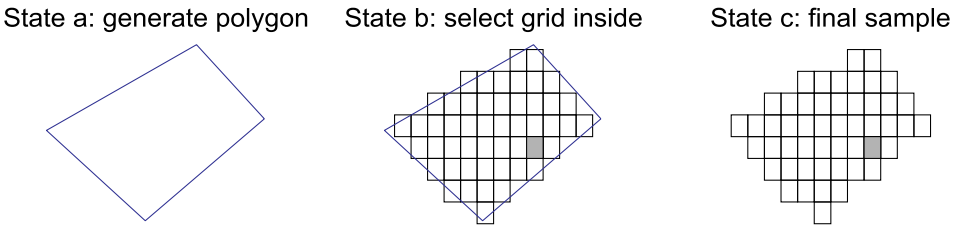

Fig. 7. The process of generating samples by shape-constrained sampling method.

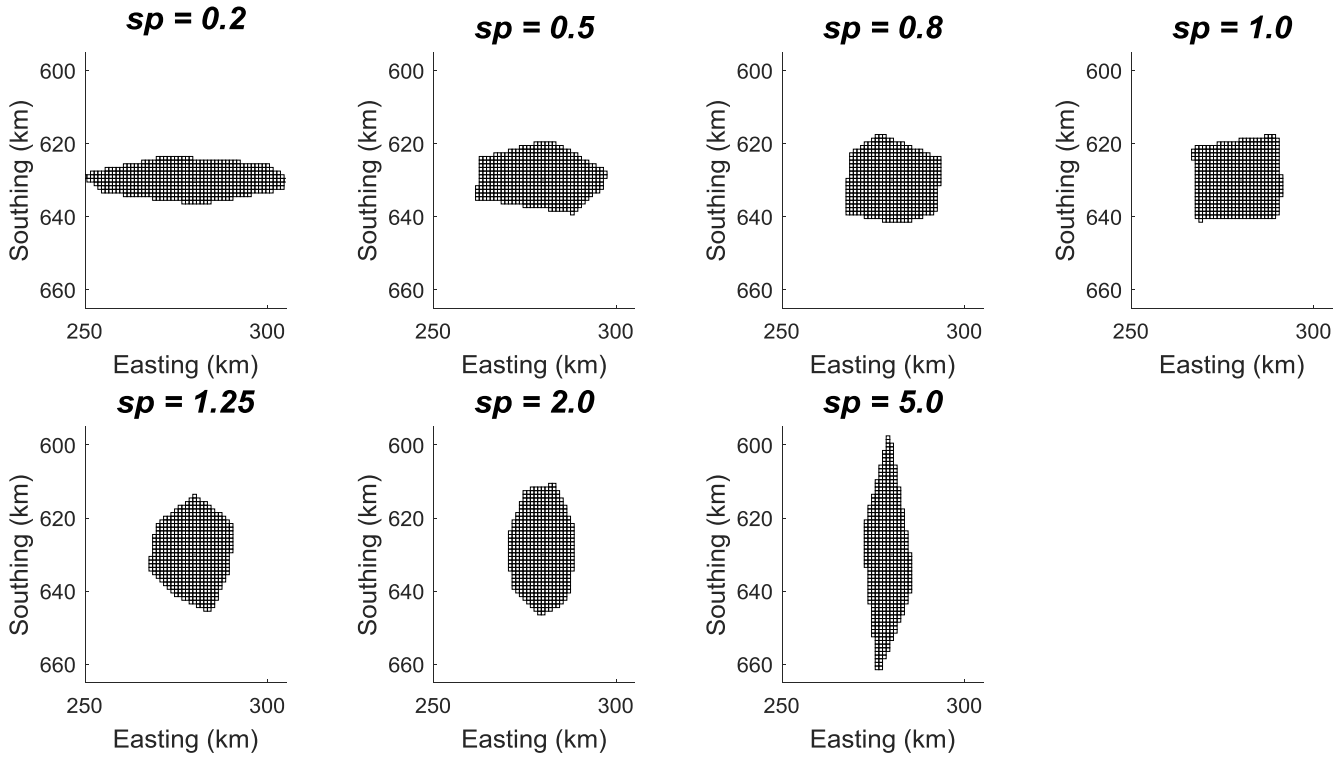

Fig. 8. The 7 samples with different spatial index sp.

south direction dominated shapes while smaller values indicate west-east direction dominated samples. Apparently, other shape related constraints can be defined and applied subject to the needs of different applications.

The value of the toolbox can well be appreciated in the analysis results, partly shown in Fig. 9, in finding the spatial variation of extreme rainfall over the GB. The entire analysis is not presented here; however, with the help of the SRS-GDA toolbox, we were able to reveal patterns never reported before. For example, a west-east variation of the rainfall distribution at different quantiles is clearly seen as "west high, east low" in Fig. 9a. What is more interesting is the symmetric pattern shown in Fig. 9b (around $sp = 1.0$) with regards to the sample shape which implies that sampled areas with slight elongation in north-south direction are expected to have a higher amount of rainfall than those spread more in east-west direction at given frequency/return period. For samples with the same size and location, there is a remarkable difference in areal averaged rainfall between more elongated (e.g. $sp = 0.2$ or 5.0) and rounded shape (e.g. $sp = 1.0$), which can be attributed to heterogeneity of the grid rainfall distribution that cannot compensate to the areal average. The relationship between the sample size and the annual maximum daily rainfall (Fig. 9c) is shown to have largely followed what is expected, e.g., decrease of areal rainfall as sample size grows.

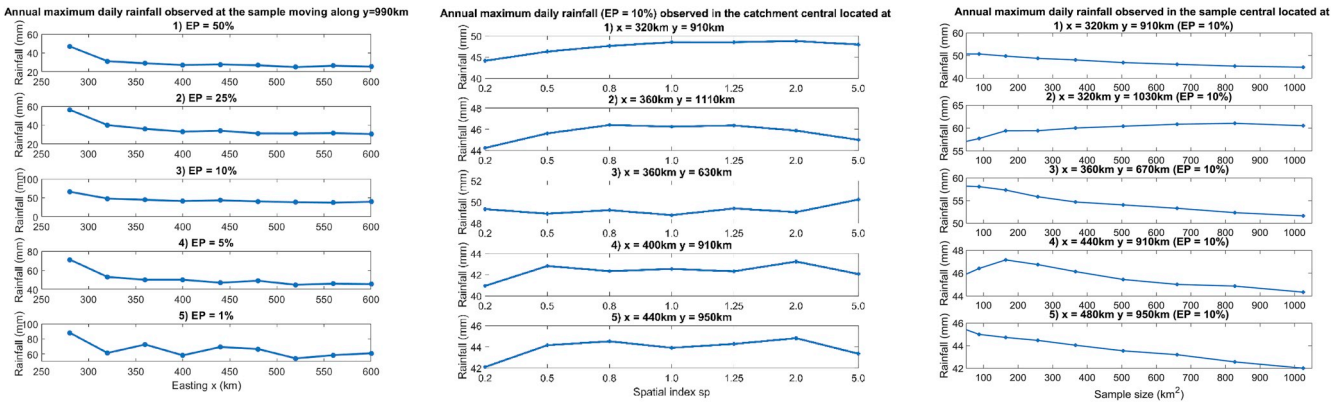

Fig. 9. The dependencies on the locations, the spatial index and the size of rainfall distribution over GB: (a) the east-west pattern and (b) the symmetric pattern with regards to the sampled shape and (c) the trend pattern with regards to the sampled area size as detected by using the toolbox discussed in this paper. EP is short for "Exceedance Probability".

4. Conclusions and availability of the toolbox

In this paper, we discuss the development of a new MATLAB toolbox for spatial random sampling in grid-based data analysis (SRS-GDA). The main aim of the toolbox is to address the very needs of many climate change related studies on spatial-temporal diagnostics of hydro-climatic datasets. An example application case is given in which the implementation details are discussed. Our initial applications show that with this toolbox, several important variation patterns of extreme rainfall (due to be published separately) over GB that have yet to be reported are clearly identified. Based on the promising results, we expect this toolbox, thanks to the availability of its source code, will help the related research community in their analyses of grid data sets and gain further insight into the underlying science.

The source code of the toolbox as well as the example case given above are available at the GitHub (https://github.com/wanghan924/SRS-GDA_Toolbox.git). The source code is provided subject to a GPL V3 licence. Use/fork of the toolbox is subject to proper acknowledgement as stated on the Webpage of the toolbox.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgements

The authors would like to thank the Centre of Hydrology and Ecology (CEH), UK for providing the GEAR dataset to test the toolbox. The co-author Han Wang's PhD study is jointly sponsored by the two scholarships offered by the Chinese Scholarship Council (CSC), China and the College of Engineering, Swansea University, UK, which are both gratefully acknowledged. This study is supported by the UK-China Urban Flooding Programme Grant (REF: UUFRIP\10021) from the Royal Academy of Engineering, United Kingdom.

References

Alamou, E.A., Obada, E., Afouda, A., 2017. Assessment of future water resources availability under climate change scenarios in the Mékrou basin, Benin. J. Hydrol. 4 (4), 51.

Bauer, P., Thorpe, A., Brunet, G., 2015. The quiet revolution of numerical weather prediction. Nature 525 (7567), 47–55.

Bell, F.C., 1976. The Areal Reduction Factor in Rainfall Frequency Estimation.

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. J. Hydrol 377 (3–4), 335–350.

Besag, J., Diggle, P.J., 1977. Simple Monte Carlo tests for spatial pattern. J. R. Stat. Soc. Ser. C Appl. Stat. 26 (3), 327–333.

Chang, K.T., 2008. Introduction to Geographic Information Systems, vol. 4. McGraw-Hill, Boston.

Chen, Y., Li, Z., Fan, Y., Wang, H., Deng, H., 2015. Progress and prospects of climate change impacts on hydrology in the arid region of northwest China. Environ. Res. 139, 11–19.

Cole, S.J., Moore, R.J., 2008. Hydrological modelling using raingauge- and radar-based estimators of areal rainfall. J. Hydrol 358 (3–4), 159–181. https://doi.org/10.1016/j.jhydrol.2008.05.025.

Covey, C., AchutaRao, K., Cubasch, U., Jones, P., Lambert, S., Mann, M., Phillips, T., Taylor, K., 2003. An overview of results from the coupled model intercomparison Project. Glob. Planet. Chang. 37 (1–2), 103–133.

Drusch, M., Wood, E.F., Gao, H., Thiele, A., 2004. Soil moisture retrieval during the Southern Great Plains Hydrology Experiment 1999: a comparison between experimental remote sensing data and operational products. Water Resour. Res. 40 (2).

Du, J.C., Zhang, Q., 2019. Spatiotemporal variability and trends in the hydrology of the Xiang River basin, China: extreme precipitation and streamflow. Arab. J. Geosci. 12 (18), 566.

Duane, W., Livingstone, D., Kidd, D., 2000. Integrating environmental models with GIS: an object-oriented approach utilising a hierarchical data format (HDF) data repository. Trans. GIS 4 (3), 263–280.

Erler, A., Frey, S., Khader, O., d'Orgeville, M., Park, Y., Hwang, H., Lapen, D., Richard Peltier, W., Sudicky, E., 2019. Simulating climate change impacts on surface water resources within a lake-affected region using regional climate projections. Water Resour. Res. 55 (1), 130–155. https://doi.org/10.1029/2018WR024381.

Fairman Jr., J.G., Schultz, D.M., Kirshbaum, D.J., Gray, S.L., Barrett, A.I., 2017. Climatology of size, shape, and intensity of precipitation features over Great Britain and Ireland. J. Hydrometeorol. 18 (6), 1595–1615.

Giorgetta, M., Jungclaus, J., Reick, C., Legutke, S., Bader, J., Böttinger, M., Brovkin, V., Crueger, T., Esch, M., Fieg, K., Glushak, K., 2013. Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5. J. Adv. Model. Earth Syst. 5 (3), 572–597.

Golding, B.W., 1998. Nimrod: a system for generating automated very short range forecasts. Meteorol. Appl. 5 (1), 1–16.

Guenzi, D., Fratianni, S., Boraso, R., Cremonini, R., 2016. CondMerg: an open source implementation in R language of conditional merging for weather radars and rain gauges observations. Earth Sci. Inform. 10 (1), 127–135.

Islam, T., Rico-Ramirez, M.A., Srivastava, P.K., Dai, Q., 2014. Non-parametric rain/no rain screening method for satellite-borne passive microwave radiometers at 19–85 GHz channels with the Random Forests algorithm. Int. J. Remote Sens. 35 (9), 3254–3267. https://doi.org/10.1080/01431161.2014.903444.

Kay, A.L., Davies, H.N., Bell, V.A., Jones, R.G., 2009. Comparison of uncertainty sources for climate change impacts: flood frequency in England. Clim. Change 92 (1–2), 41–63.

Monteiro, J.A., Strauch, M., Srinivasan, R., Abbaspour, K., Gücker, B., 2016. Accuracy of grid precipitation data for Brazil: application in river discharge modelling of the Tocantins catchment. Hydrol. Process. 30 (9), 1419–1430.

Moufouma-Okia, W., Jones, R., 2015. Resolution dependence in simulating the African hydroclimate with the HadGEM3-RA regional climate model. Clim. Dyn. 44 (3–4), 609–632.

NERC, 1977. Flood Studies Supplementary Report No 1: the Areal Reduction Factor in Rainfall Frequency Estimation. Natural Environment Research Council, UK.

Ning, S., Wang, J., Jin, J., Ishidaira, H., 2016. Assessment of the latest GPM-era high-resolution satellite precipitation products by comparison with observation gauge data over the Chinese Mainland. Water 8 (11), 481. https://doi.org/10.3390/w8110481.

Ordnance Survey, 1946. A Brief Description of the National Grid and Reference System. His Majesty's Stationery Office (HMSO), London.
Rabiei, E., Haberlandt, U., 2015. Applying bias correction for merging rain gauge and radar data. J. Hydrol 522, 544–557.
Rew, R., Davis, G., 1990. NetCDF: an interface for scientific data access. IEEE Comput. Graph. Appl. 10 (4), 76–82.
Rojas, R., Feyen, L., Dosio, A., Bavera, D., 2011. Improving pan-European hydrological simulation of extreme events through statistical bias correction of RCM-driven climate simulations. Hydrol. Earth Syst. Sci. 15 (8).
Rutledge, G.K., Alpert, J., Ebisuzaki, W., 2006. NOMADS: a climate and weather model archive at the National Oceanic and Atmospheric Administration. Bull. Am. Meteorol. Soc. 87 (3), 327–342.
Schaake, J.C., Hamill, T.M., Buizza, R., Clark, M., 2007. HEPEX: the hydrological ensemble prediction experiment. Bull. Am. Meteorol. Soc. 88 (10), 1541–1548.
Silver, M., Karnieli, A., Marra, F., Fredj, E., 2019. An evaluation of weather radar adjustment algorithms using synthetic data. J. Hydrol 576, 408–421. https://doi.org/10.1016/j.jhydrol.2019.06.064.
Sinclair, S., Pegram, G., 2005. Combining radar and rain gauge rainfall estimates using conditional merging. Atmos. Sci. Lett. 6 (1), 19–22.
Skamarock, W.C., Klemp, J.B., Dudhia, J., 2001. Prototypes for the WRF (weather research and forecasting) model. In: Preprints, Ninth Conf. Mesoscale Processes, J11–J15. Amer. Meteorol. Soc., Fort Lauderdale, FL.
Smith, R.C., Cheeseman, P., 1986. On the representation and estimation of spatial uncertainty. Int. J. Robot. Res. 5 (4), 56–68.
Svensson, C., Rakhecha, P.R., 1998. Estimation of probable maximum precipitation for dams in the Hongru River catchment, China. Theor. Appl. Climatol. 59 (1–2), 79–91.
Tanguy, M., Dixon, H., Prosdocimi, I., Morris, D.G., Keller, V.D., 2016. Gridded Estimates of Daily and Monthly Areal Rainfall for the United Kingdom (1890-2009) [CEH-GEAR]. NERC Environmental Information Data Centre. Retrieved from NERC Environmental Information Data Centre. https://doi.org/10.5285/33604ea.
Thorndahl, S., Einfalt, T., Willems, P., Nielsen, J.E., ten Veldhuis, M.C., Arnbjerg-Nielsen, K., Rasmussen, M.R., Molnar, P., 2017. Weather radar rainfall data in urban hydrology. Hydrol. Earth Syst. Sci. 21 (3), 1359–1380.
Uppala, S.M., Kållberg, P.W., Simmons, A.J., Andrae, U., Bechtold, V. Da Costa, Fiorino, M., Gibson, J.K., Haseler, J., Hernandez, A., Kelly, G.A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R.P., Andersson, E., Arpe, K., Balmaseda, M.A., Beljaars, A.C.M., Berg, L. Van De, Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B.J., Isaksen, L., Janssen, P.A.E.M., Jenne, R., McNally, A.P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N.A., Saunders, R.W., Simon, P., Sterl, A., Trenberth, K.E., Untch, A., Vasiljevic, D., Viterbo, P., Woollen, J., 2005. The ERA-40 re-analysis. Q. J. R. Meteorol. Soc. 131, 2961–3012.
Viviroli, D., Weingartner, R., Messerli, B., 2003. Assessing the hydrological significance of the world's mountains. Mt. Res. Dev. 23 (1), 32–41.
Weather Bureau, U.S., 1958. Rainfall Intensity-Frequency Regime Parts 1 and 2, Technical Paper No. 29. Department of Commerce, U.S, Washington D.C., US.
Wu, S., Angelikopoulos, P., Papadimitriou, C., Koumoutsakos, P., 2018. Bayesian annealed sequential importance sampling: an unbiased version of transitional Markov chain Monte Carlo. ASCE-ASME J. Risk Uncertain. Eng. Syst. Part B Mech. Eng. 4 (1), 011008.
Xu, C., He, H.S., Hu, Y., Chang, Y.L., Bu, R., 2005. Latin hypercube sampling and geostatistical modeling of spatial uncertainty in a spatially explicit forest landscape model simulation. Ecol. Model. 158 (2–4), 255–269.