Global flood hazard mapping using statistical peak flow estimates

C. Herold[1] and F. Mouton[2]

[1]{UNEP/GRID-Europe, Geneva, Switzerland}

[2]{UFR de Mathématiques, Université J. Fourier, Grenoble, France}

Correspondence to: C. Herold (christian.herold@unepgrid.ch)

Abstract

Our aim is to produce a world map of flooded areas for a 100 year return period, using a method based on large rivers peak flow estimates derived from mean monthly discharge time-series. Therefore, the map is supposed to represent flooding that affects large river floodplains, but not events triggered by specific conditions like coastal or flash flooding for instance.

We first generate for each basin a set of hydromorphometric, land cover and climatic variables. In case of an available discharge record station at the basin outlet, we base the hundred year peak flow estimate on the corresponding time-series. Peak flow magnitude for basin outlets without gauging stations is estimated by statistical means, performing several regressions on the basin variables. These peak flow estimates enable the computation of corresponding flooded areas using hydrologic GIS processing on digital elevation model.

1 Introduction

Recent developments and reports on global risk identification (Disaster Risk Index (DRI)), the central component of the report "Reducing Disaster Risk" by the United Nations Development Programme (UNDP/BCPR, 2004 and Peduzzi et al., 2009) and World Bank "Disaster Risk Hotspots" (Dilley et al., 2005) lead, in particular, to conclusions about precision of the generated natural hazard maps. Considering flood hazard, applied methodologies clearly demonstrated a lower resolution in their resulting global map, compared to other hazards. Basically, the two developed hazard maps highlight basins prone to flooding more than they delimit zones potentially at risk. Consequently, it was underlined that new developments regarding flood hazard map would be essential during further efforts in global risk identification, in order to obtain a similar level of spatial resolution among the different natural hazards.

Motivated by the Development Research Group at the World Bank, a preliminary study was achieved in order to demonstrate that a specific methodology was applicable at a global scale to produce a relevant global flood map (Herold and Mouton, 2006). As far as choice of methodology is concerned, spatial compilation of existing recorded flood events could not be considered, since existing global databases would not present a satisfactory spatial coverage. Following a suggestion of K.L. and J.P. Verdin at EROS Data Center (EROS/USGS), we made the choice to test a statistical method known as "Peak Flow Estimates" (see for instance Sando, 1998). The main idea was to estimate, for each basin of a certain size, the flooded area corresponding to a hundred year recurrence peak flow, using the peak flow discharge and a digital elevation model (DEM). For basins with a gauged station close to their outlet, this discharge could be estimated by statistical modeling using the time series of annual peak flows. For basins without gauged station, it could be estimated by regression formulae established on groups of gauged basins with similar hydromorphometric, land cover and climatic values.

Regarding statistical methods, there was a need for simplicity and robustness as they had to be applied to a huge amount of data and to be automated as far as possible. Concerning the hundred year peak flow estimated from annual time series, we had to choose a statistical model as well as a procedure for the estimation of parameters. Quoting Mkhandi et al. (2000): "…it is not possible to identify a parent distribution for the annual maximum floods. Attempts over many years have proved inconclusive." However, two main distributions are used in

practice for that purpose: the generalized extreme value one (GEV) and the log-Pearson III one. Both have proved to give acceptable results, the better one depending on the specific case. Although local studies require a fine choice of distribution (Meigh et al., 1997), in the case of a global study, we were forced to use the same distribution for all basins. We decided to follow the methodology prescribed in the bulletin 17b of the United States Water Resources Council's Hydrology Subcommittee (1982), which use log-Pearson III. For estimation of the parameters, there are several methods as well, and comparison is far from straightforward (see Dupuis, 1999, in the case of GEV). Here, we used the method of moments following bulletin 17b. That combination of model and estimation has proved to be relevant in several cases (see for instance Mkhandi et al., 2000). Concerning the regressions, we used the simplest method, i.e. linear models after suitable transformations of variables.

During the preliminary study (Herold and Mouton, 2006), the method was tested on specific sub-basins of North and South American continents. As a conclusion, the study showed that, under specific conditions, the methodology would probably be applicable at a global scale and give satisfactory results.

Later on, decision was taken to apply this methodology at a global scale, as a part of the ISDR system's new effort on Global Risk Identification. The expected final product is a global probabilistic map of flooded areas for a hundred year return period, using the Peak Flow Estimates methodology along with required global datasets. As this method is based on large river discharge time-series, it is supposed to represent events that affect corresponding floodplains. The model is not expected to properly represent events triggered in different conditions, for instance coastal or flash flooding. The final map has to give satisfactory results in the case of this newly undertaken global risk analysis. It will not provide the level of precision required for local analysis or land use planning. The applied methodology is the one developed in the preliminary study, except for certain points that will be detailed in the text. For instance, new datasets were available. We also use weighted linear models for regressions instead of simple linear models, which proved to be comparable to general linear models (Kjeldsen et al., 2001).

Flooded zones are generated using a model provided by EROS Data Center (EROS/USGS). Results are compared to a 10 year record of flood events provided by Dartmouth Flood Observatory (DFO). In general, flooded zones generated by the model tend to be smaller than footprints available through DFO database, in particular for floodplain having very large

drainage area. In order to benefit from advantages of both sources, the final map is obtained by using both events recorded in the DFO database and modeled flooded areas.

The organization of the paper is the following: the datasets are presented in section 2, the methodology in section 3, the results in section 4 and the discussion in section 5.

Notice that in a similar effort to enhance a global approach of flood hazard, K.L. and J.P. Verdin are leading a five-year project called “Development and Implementation of Globally Applicable Methods for Characterization of Flood Hazards”, which aims to develop a methodology in collaboration with local experts. That project has been funded through the U.S. Office of Foreign Disaster Assistance.

Acknowledgements

This work was supported by the International Strategy for Disaster Reduction (ISDR).
We would like to thank the following persons who have collaborated to this project:

James P. Verdin, Kristine L. Verdin., Kwabena Asante, G. Robert Brakenridge, E. Anderson, Uwe Deichmann, and Thomas Moelhave.

2 Data

2.1 River discharge datasets

The river station dataset is composed of georeferenced stations and their recorded mean monthly discharge time series. It is a compilation of global, regional and national datasets collected by various research centers. As the main aim is to reach an acceptable global coverage, special effort is made to access data provided by national services whenever possible.

2.2 Digital Elevation Model and derived hydrological datasets

Three Digital Elevation Model are used during various stages of the project:

HYDRO1K (EROS, USGS) is used for generation of a first set of variables for statistical analysis. The one kilometer resolution and the availability of ancilliary products of this dataset are considered as most relevant for this stage of analysis.

Global Drainage Basin Database (GDBD) is used in some specific cases to help correcting HYDRO1K modeled river network.

HydroSHEDS (WWF. In partnership with USGS, CIAT, TNC, CESR) is used to calculate peak flow estimates and generate corresponding flooded areas. The 90 meters resolution of this dataset is considered as essential in the process of generating flooded zone patterns.

2.3 Climatic datasets

Monthly precipitation and monthly mean temperatures global raster provided by the Climatic Research Unit at University of East Anglia are used to generate two variables: mean annual precipitation and minimum mean monthly temperature. For the purpose of this study, these two datasets show relevant spatial resolution and time extent.

Variability Analyses of Surface Climate Observations (VASClimO) provided by the Global Precipitation Climatology Centre (GPCC) is used to generate the variable Monthly maximum precipitation for a 100-year return period. This dataset is chosen for its reliability and homogeneity in time.

Three different climate classification maps are used to associate each basin to corresponding climatic zones and help grouping basins during further statistical analysis. The first dataset,

the Holdridge Life Zones data set, was already used in the preliminary study. It is completed with two recently available map of Köppen-Geiger climate classification at two different resolutions:

- The World Map of the Köppen-Geiger climate classification updated (Kottek et al., 2006);
- The Updated world map of the Köppen-Geiger climate classification (Peel et al., 2007).

2.4 Land cover datasets

Global land cover GLC_2000 version 1 (IES Global Environment Monitoring Unit) is used to generate two different variables: Forest cover and Impervious cover.

The Global Lakes and Wetlands Database (GLWD) is used to generate the Surface water storage variable.

Both datasets present adequate precision for generating these three variables.

2.5 Recorded flood event dataset

Flood event inundated areas recorded in the World Atlas of Flooded Lands and provided by Dartmouth Flood Observatory are used to validate the final pattern of flooded zones generated by the model. As the only global database at such time extent and spatial resolution, it is considered as essential for this study.

3 Methodology

The global process is represented on the flow chart of Figure 1. It follows three steps: the production of basin variables by a first spatial analysis, the production of groups and peak flow models by statistical analysis and the estimation of peak flows and flooded areas by a second spatial analysis.

3.1 Methodology for the production of basin variables

The process used for generating a set of variables, suitable for the regression analysis, is represented on the flow chart of Figure 2. The production of the dependent variable, represented by a set of selected georeferenced gauging stations that match basin outlets and corresponding time-series of monthly mean discharge, is the delicate part of the spatial analysis, as detailed below. The list of 16 independent variables (Table 1) is inspired by Verdin (personal communication) and (Sando, 1998), within the limitations of available global datasets. It can be classified in three categories: hydromorphometric, land cover and climatic. These independent variables are generated for the drainage basin of each selected gauging station, by classical G.I.S. techniques. Most of the procedures described in this section are automated using ArcInfo Macro Language (AML).

3.1.1 Basins

To minimize usage of time-consuming GIS procedures, and considering a resolution of 1 kilometer as satisfactory, we decide to base on the HYDRO1k dataset the process of generating variables for the statistical analysis. As it is partly based on drainage area, this process is applied separately on each HYDRO1k sub-region, in order to maintain Lambert azimuthal equal-area projection. Hydrologic corrections of HYDRO1k DEM are made if necessary.

In order to structure spatial analysis, and to avoid too much dispersion in basin areas during the statistical analysis, we have to consider HYDRO1k basin outlets of a specific Pfafstetter hierarchical level as a spatial reference. The system developed by Otto Pfafstetter is based upon the topology of the drainage network and the size of the drained surface area. Its numbering scheme is self-replicating, making it possible to provide identification numbers to the level of the smallest sub-basins extractable from a DEM (Verdin, 1997). We select level 5 of Pfafstetter code, as it presents an appropriate balance between basin area and spatial

density of outlets. This density has a direct influence on the number of total available stations that is maintained in the final dataset after treatment, especially in regions with a low density of gauging stations.

3.1.2 Selection and adjustment of gauging stations

A delicate point is the selection and the spatial adjustment of the river discharge stations that can reasonably match a basin outlet. From each available discharge station dataset, a subset is selected including stations with at least 1000 km^2 of drainage area, and a minimum record of 7 years with 12 monthly means. Then, each of these subsets is formatted and integrated in a unique database. When the same station is present in two different datasets, precedence is first given to dataset for which clear information on monthly mean calculation is available, then to stations with the maximum available years of records.

Because of HYDRO1k resolution and possible imprecision in station recorded information (drainage area and geographic coordinates), spatial adjustment is required between the two datasets. The first process moves each station to the closest HYDRO1k stream section. At that point, any station such that the difference between recorded and HYDRO1k drainage area is below 10%, is considered to be adequately located on the stream network. Other stations are moved up- or downstream until the same threshold of area difference is reached.

Then, each of the available stations is moved again to the nearest level 5 outlet. In order to affect a maximum number of stations to these outlets, respective drainage areas are considered. For each station, the nearest outlet downstream is considered, as well as the nearest one upstream. If the area of the station represents at least 75% (resp. at most 150%) of the downstream (resp. upstream) outlet area, the station is moved to that outlet. Accordingly, their recorded discharge values are divided by the same area ratio (recorded/ HYDRO1k). During this process, if more than one station verifies the condition on drainage areas in the up- or downstream basin of a level 5 outlet, the selection is made considering first the available years of records, and then the area ratio.

In order to avoid spatial redundancy that might affect statistical analysis, the final subset excludes any station the drainage basin of which includes the basin of another station. This means that any station downstream of another station is not included in the final dataset. The Pfafstetter code assigned to each HYDRO1k stream section is an efficient key to automatically perform this selection. At this stage, a station dataset is generated, including,

for each basin outlet, a unique station code, level 5 Pfafstetter code, and available discharge records.

3.2 Methodology for the production of peak flow estimation models

As shown on the flow chart of Figure 3, the statistical analysis consists of two phases: the first one produces the statistical variables and the second one produces groups and regression models, which enable peak flow estimates for ungauged sites. The methodology follows the directions of the Bulletin 17B from United States Water Resources Council's Hydrology Subcommittee (USWRC, 1982) and (Sando, 1998).

Certain parts of this process are easily automated by way of programming, but human interpretation is necessary for some crucial steps, namely the grouping of basins and the choice of the "best" regression formulae, even with the help of statistical software.

3.2.1 Statistical variables

Peak flow corresponding to a hundred year recurrence interval are estimated following (USWRC, 1982): an acceptable modeling of the distribution of the observed annual peak flows for a given site is the log-Pearson type III law, which involves three parameters: the mean μ, standard deviation σ and skew coefficient G of the log of peak flows. These parameters are estimated by the method of moments, and the formulae are easily calculated from the series of observations. After standardization (subtracting the mean and dividing by standard deviation), we compute the inverse cumulative density function of standard Pearson type III law with the same skewness, for the probability corresponding to the recurrence interval (e.g. 0.99 for a 100 year recurrence interval). Since there is no exact formula and the skew coefficients of different stations are different (which prevents from reading the result in a table), we use the approximate formula given in (USWRC, 1982):

$$K = \frac{2}{G}\left[\left(\left(K_n - \frac{G}{6}\right)\frac{G}{6} + 1\right)^3 - 1\right]$$

where K is the value of the inverse cumulative probability function for the above probability, G is the skew coefficient and K_n is the standard normal deviate corresponding to the same probability. Note that this approximation is good for G between –1 and 1, which should be the case for most of the stations (see the exploratory study for North and South America and the

map of (USWRC, 1982) in the case of North America). The log of the peak flow estimation is then given by

$$\log(Q) = \mu + \sigma K.$$

A variable of exceptional precipitation, corresponding to the same recurrence interval, is obtained by the same method.

Furthermore, most of the variables need to be transformed using the logarithm in order to take into account non-linearity in the regression (see Sando, 1998) and also particular distributions of initial variables.

All these operations are easily automated.

3.2.2 Groups and regressions

A 1-variable analysis of the statistical variables is performed in order to check for particularities of their distributions. The links between the variables are studied and possibly explained (for instance by physical reasons).

In order to compute regression formulae, it is necessary to constitute groups of stations, which are homogeneous from the point of view of basin, climatic and geographic characteristics. Nevertheless, the choice between different possible groupings depends on the quality of the regressions performed on the different groups.

Once the "best" grouping is fixed, we choose the "best" regression formula for each group, estimating peak flows given basin and climatic variables.

3.3 Methodology for the estimation of flooded areas

The process, based on the HydroSHEDS DEM dataset, is described on the flow chart of **Figure 4**. It consists of generating flooded area in each basin using peak flow estimates along with hydrological model based on Manning's equation. Because of the HydroSHEDS 90-meter resolution, this process is applied separately on 46 groups of basins. Variables selected by the regressions (**Table 5**) are generated at each stream section outlet in order to calculate the required peak flow estimates.

3.3.1 HydroSHEDS Digital Elevation Model

In order to apply the hydraulic model generating flooded areas on HydroSHEDS conditioned elevation, we have to generate a range of ancillary products from the available tiles, required to compute the variables.

First, individual basins as delimited by HydroSHEDS shapefiles, are grouped into 46 sets, each of them being totally included in one of the six HYDRO1k datasets. For each group, we merge the corresponding tiles and clip them to the basin group boundaries. At this point, some pixels with no value are identified in some original tiles. They are considered as errors and set to the minimum value of the 8 pixel direct neighborhood during the procedure. Then, we produce flow direction and accumulation grids using the *ts-route* and *ts-accumulate* executables of TerraSTREAM stand-alone application (Danner et al., 2007). A threshold of 1000 km^2 is applied on flow accumulation raster to produce the stream network. Then, the "streamlink", "watershed" and "streamline" functions are applied in order to generate stream sections as lines with unique ID and correspondent watershed polygons.

We generate only the seven variables that are significant in the regressions (Table 5). They are computed at each HydroSHEDS stream section outlet, but based on a 1 km^2 resolution. For the sake of consistency, the two variables based on altitude (Mean basin elevation and Mean basin slope) are generated using the HYDRO1k DEM, as those used in the regressions computations are.

At this stage, the procedure of moving stations along stream network (described in section 3.1.2) is reapplied on the HydroSHEDS dataset. Therefore, spatial correspondence is established between gauging stations and basin outlets. It allows comparison between peak flow estimates derived from station time series using the log-Pearson type III distribution, and those based on regression equations.

In order to accelerate data processing, the procedure described in this section is automated in one single ArcInfo Macro Language code (AML), calling executables and ArcGIS Visual Basic code when needed.

3.3.2 Model generating flooded area

Flooded areas are generated using a hydraulic model provided by the EROS Data Center (EROS/USGS). The model first generates a relative DEM from HydroSHEDS that set any

stream pixel values to 0 as a reference altitude. Then, it generates cross sections of a specified width for each stream section. Each cross section is used to extract altitude values from relative DEM and generate a specific stage vs. discharge function using Manning's equation. These functions are finally used to calculate river stage from peak flow estimates for a 100 year recurrence interval, and then generate corresponding flooded areas for each stream section basin, using the generated relative DEM.

For the specific case of this project, we add a procedure to the model. Its function is to automatically adjust the cross sections orientation considering mean azimuth of the corresponding stream section in a 1 km radius.

4 Results

4.1 Production of basin variables

The variables are produced according to the methodology. Two points of interest are detailed here: the hydrological correction of the DEM and the density of gauging stations.

4.1.1 Correction of HYDRO1k Digital Elevation Model

The Global Drainage Basin Database is used as a reference in some of the modifications described here below:

- Corrections are applied to the European HYDRO1k stream network. In particular, it modifies the source of the Rhone and, as a consequence, shortens a confluence of Rhine.
- Some outlet stream sections on the Atlantic Ocean, Mediterranean and Caspian seas are modified to correspond spatially to basin layer, in order to have subsequent processes running correctly.
- As the Australian stream network is missing Pfafstetter code, we develop an automatic procedure to rebuild this information.

4.1.2 Discharge stations

The station dataset is mainly composed of global and regional compilation of data, available online or under specific request to the official provider. Here is a short description of the coverage:

- In North America and Australia the dataset coverage is very good.
- In the case of Europe, it is possible to collect information from some national provider when needed, in order to complete global datasets. We finally add national datasets from France, Portugal, Spain and Switzerland.
- Because of the relatively poor density of gauging stations in some regions, such as South America, Asia and Africa, we end up trying to base the statistical analysis on a global approach as detailed in subsection 4.2.2**.**

The effect of the successive selections (described in the methodology) on the station dataset can be seen on Table 6, Figure 5, Figure 6, Figure 7 and Figure 8, which show the distribution of collected stations at different stages of treatment.

4.2 Production of peak flow estimation models

4.2.1 Statistical variables

According to the methodology, logarithms of exceptional peak flows are easily estimated in an algorithmic way, which gives the variable LQ100. In the same way, we also produce a variable called LogP100, which is the logarithm of the hundred year exceptional monthly precipitation.

The next step is to transform the variables according to the preliminary study guidelines (Herold and Mouton, 2006). The log-transformed variables are denoted LDRAREA, LMEANALT, LMNSLOP, LKGRAV, LDRFREQ, LSOIL_HC, LMCHLENGTH, LMCHSLOPE and LPRMEAN. The variable FORCOV, being a percentage, needs to be transformed before taking its logarithm, in order to avoid getting only negative values: we take the logarithm of T(FORCOV), where T(x)=x/(1-x), noted LTFORCOV. Since the variables WATER_STOR, URBCOV often take zero values, they do not allow log-transformation and hence are not taken into account in the regression. The variable CLDERMONTH already takes range in negative and positive values and there is no physical reason to justify a transformation (which would enable to take the logarithm but would be artificial). In addition to the Holdridge climatic variable used in the preliminary study and deduced from the Holdridge Life Zones classification, two new variables are constructed, using two different recent studies based on the Köppen-Geiger classification (Peel et al., 2007) and (Kottek et al., 2006); these variables are labeled Koge1 and Koge5, and are obtained by considering for each basin, the climate class obtaining the maximum area.

4.2.2 Groups and regressions

Descriptive statistics

Matrix plots show a strong correlation between LDRAREA and LMCHLGTH, and correlations between the three variables CLDMONTH, LPRMEAN and LogP100 on one hand and between the three variables LMNSLOP, LMEANALT and LMCHSLOP on the

other. A PCA and its circle of correlation confirm the existence of those three groups, in each of which at most one variable has to be selected as an independent variable in the regressions.

Constitution of groups

The first try is a regional one, for two reasons. Firstly, the preliminary study (Herold and Mouton, 2006) showed that regional regressions could give some good results. Secondly, as the process of acquisition and treatment of data at a global scale is still not achieved at this stage (beginning of the project), such a regional treatment allows to begin the statistical analysis sooner. It gives some quite good results and enables to compare the three climatic variables. For the Holdridge classification, in the preliminary study we used seven groups of classes for North America and three for South America. We use here the same seven groups for North America, Asia and Europe and the same three groups for South America, Africa and Australia. For the two Köppen-Geiger classifications we use the five groups A, B, C, D and E (See K.-G. classification in Table 2). The three climatic zone variables give some similar results, but slightly better for Köppen-Geiger classifications (variables Koge1 and Koge5). In general, the regressions by groups are good, some are excellent and a few are poor.

At this point, we have to choose between two strategies: refining those regional regressions to obtain some acceptable results in all cases or trying a global approach because the global data is available then. We choose to try the second one: a global approach is certainly ambitious but if it gives some results, it would certainly be more robust and would compensate the lack of global homogeneity of the data to some extent.

Global Approach

Climatic groups are established at the global level. Firstly, we use the seven groups used above in North America for Holdridge variable and the five groups for the Koge1 and Koge5 variables. This rough study shows that there is an issue for the regressions on the groups concerning deserts and steppes, for all three climatic variables.

Secondly, we have to refine this rough grouping. Because of the previous remark on comparison between the three climatic variables, we focus on the two variables Koge1 and Koge5. Moreover, these two variables are easily compared, since they use the same theoretic classification. For this, we use the first two letters of the Köppen-Geiger classification (Table 2), i.e. without the third letter (a, b, c, d, h or k); this gives 12 groups, from Af to E (class ET and EF are merged due to their small size). Groups constructed according to Koge1 give slightly better regressions, so we choose that variable for the final process.

Thirdly, we have to refine those groups. This is the hardest part, requiring lots of trials. For three of these refined groups, we find that an additional regional subdivision is necessary, and possible according to their size. Here, we only give the final grouping that uses variable Koge1 (Table 3).

Regressions

For each group, a "best subset" regression is performed to have a general picture. Then, combining contradictory arguments such as better R-square and significance of variables, with help of Mallow's Cp and a systematic analysis of residuals, searching for a very small number of variables in the case of small groups (and a limited number of variables for the others), adding or subtracting variables one by one of the model, we select a "best" regression formula. For some groups, one or more outliers have been taken apart for establishing regression formula. For most of the groups, results are very satisfactory. For a few, they are less significant, and for one (group 6, hot desert), no regression is possible. The results are summarized in Table 4.

Remark

We also estimate peak flows corresponding to a 50 year recurrence interval, keeping the same groups and reprocessing the regressions. The significant variables are with no doubt the same as in the 100 year case, except for one or two groups, where there is a close competition between two sets of variables.

4.3 Estimation of flooded areas

4.3.1 HydroSHEDS Digital Elevation Model

The TerraSTREAM application (Danner et al., 2007) is essential to accumulation computations. Basically, it can process DEM over 300 million pixels within a day. For example, it runs flow direction and accumulation functions on Amazonian basin (700 million pixels) in less than ten hours, where ArcInfo or ArcGIS would run for days.

We use 5 workstations (RAM: 2-3 GB / CPU: 3.2-3.6 GHz) during about 16 days to run the process described in subsection 3.3.1 on the 46 basin groups. This represents around 1920 hours of computation.

4.3.2 Model generating flooded area

Using the same workstations as described above, we need about 3 weeks to run this process on the 46 basin groups. This represents around 2520 hours of computation.

For the following reasons, in some specific areas, inundation patterns are missing or doubtful:

- **Zone A:** Basins contained in Köppen-Geiger climatic zone called BWk (Arid-Desert-Cold), corresponding to statistical analysis group 7, show doubtful results in some instances. Apparently, it is mostly the case when discharge station network is of low density.
- **Zone B:** Flooded areas are generated using HydroSHEDS dataset, which is derived from SRTM digital elevation model. As SRTM spatial coverage includes latitudes from 60-degree north to 56-degree south, the model does not process any watershed that is beyond these limits.
- **Zone C:** A 1000 km^2 minimum threshold is applied on drainage area when generating stream network from HydroSHEDS conditioned elevation. As a consequence, any closed or coastal basin with drainage area smaller than 1000 km2 is not represented and no flooded area is generated for them.
- **Zone D:** It is not possible to find a regression for group 6, one of the groups defined during regression analysis. It corresponds to Köppen-Geiger climatic zone called BWh, described as Arid-Desert-Hot. Hence, there is no peak flow estimates for basins located in this climatic zones.

A distribution of those four zones by countries is given in Table 7. A global geographic distribution of final results and zones with no or doubtful results, according to above description, is shown on Figure 11.

5 Discussion

5.1 Discharge stations

With much time and effort, the discharge station dataset could be improved in at least two directions. Firstly, there are other European countries offering data distribution facilities, which would improve the European covering, that is however not so poor. Secondly, for South America, Asia and Africa, the dataset would have appreciably gained from contribution from some national providers. It is very difficult to obtain such information, moreover in a reasonable time frame, without personal contact (Data from Sri Lanka are obtained that way in the present study).

5.2 Peak flows: estimated vs stations

Adjusting gauging stations on HydroSHEDS stream network allows comparison between peak flow estimates derived from station time series using log-Pearson type III distribution, and those based on regression equations. Some regressions tend to overestimate peak flow (Gange, Figure 9), when other lead to underestimation (Yangtze, Figure 10). In general, regressions are more robust for drainage area smaller than 500'000 km^2, which is probably due to the fact that regression analysis samples include basins up to 250'000 km^2.

5.3 Assessing limitation of the model using detected flood events

Validation of flooded areas generated by the model is made using a 10 year record of flood events provided by Dartmouth Flood Observatory (DFO). This consists of flood events as detected by satellite sensors. Major differences between the two compared datasets are identified in specific cases described below:

- Near the cost lines where surge effect has the greatest influence. As the model is not supposed to take into account the phenomenon of coastal flooding, such events are not properly represented in the final map (see Figure 12).
- As the model generates only confluences, braided streams and corresponding basins in large floodplains are not correctly represented. In some cases, this can generate underestimation of flooded areas.

- When estimated water height is such that it would generate a theoretical overflow into neighboring basins, the model does not take it into account. This phenomenon happens mostly in the case of large floodplains and generates underestimation of flooded areas.
- As explained in subsection 4.3.2, basins contained in Köppen-Geiger climatic zone called BWk (Arid-Desert-Cold), show doubtful results in specific cases.

There are some issues that can also have an importance at specific stages of the methodology and influence final results. They should be taken into account in further development of this approach:

- Procedure to adjust the river cross sections on the stream section.
- Basins having HydroSHEDS DEM with important stream burning.
- Region presenting intense forest cover.
- Fragmentation of river network and flow by dams and reservoirs.
- Basin having specific soil and/or geology.
- Ideally, daily discharge time-series could be used in place of monthly values. But collecting such a global station datasets with relevant time and space coverage would be a very time consuming task that should not be underestimated.

Nevertheless, our results illustrate a first attempt to generate global flooded areas for a 100 year return period, whereas previous global approach could merely highlight basins prone to inundations. Combined with compiled DFO dataset, the final map shows flood patterns produced by both sources, the model and the event database. This final map gave good results during subsequent global risk analysis. The results can be accessed through the web based geoportal called PREVIEW Global Risk Data Platform (Giuliani and Peduzzi, in prep.), at the following address: http://preview.grid.unep.ch/.

5.4 Ganges and Brahmaputra

In a further effort, we modified the procedure generating flooded area and tested it on the Ganges and Brahmaputra basin. The aim is to solve two of the issues previously described: better adjust cross sections on river channel, and include neighbor basins when estimated flood height is larger than basin relative height. We choose the case of these basins because

they have an excellent coverage of recorded hundred-year events. Hence, the comparison of the model results with observed flooded areas is optimal.

Figure 13 shows the total area affected by ten years of observed events and flooded area generated by the model. It underlines that the two datasets show the best spatial fit in intermediate basins, whereas the model tends to overestimate flooded surfaces in downstream large floodplains. Figure 14 and Figure 15 highlight this tendency along Brahmaputra River. For each section of 50 kilometers, they show maximum extent of flood pattern calculated from the river channel, and total flooded area, respectively.

Figure 16 shows, in each Pfafstetter level 4 basin, the proportion of total observed flooded area that is not covered by the model. If this figure confirms that the modelized flood surfaces cover a large majority of the observed flood total area in intermediate and downstream basins, it also highlights spatial discrepancies in smaller upstream basins, particularly in the Southern part of the Ganges catchment area. This could be explained both by a lack of accuracy of our model, and by a possible remote sensing bias in the case of shallow floods. Furthermore, regarding the recorded event dataset, lateral shift due to small residual image distortion has more consequences when considering smaller flood surfaces. Anyway, finding spatial correspondence between the two datasets is not obvious regarding flood surfaces of that relatively small size.

Flood height: Analysis of model vs. observed events

Figure 17 shows, for each stream section basin, the ratio of model to observed flood height. These variables are generated using a spatial average of the relative DEM in the flooded area of each stream section basin. Observed patterns confirm the trend visible in Figure 16. Again, smaller upper basins show underestimation of flood height by the model, which could be explain by the reasons previously invoked. Best fits between the model and the observed floods are seen in intermediate and downstream basins. Plot of Figure 18 also illustrates this tendency.

To further describe the relation between modelized and observed variables, we realize a short statistical analysis. Hereafter, the variable MFL refers to the model flood height, and the variable DFO to the observed flood height.

First the 1 variable analysis, using histograms and boxplots, shows that the two distributions, MFL and DFO, have similar shapes except in a small neighborhood of zero. This reflects the

issue concerning small flood areas described before and illustrated by Figure 16 and Figure 17. Anyway, we fix a threshold on MFL for quantitative comparison. We also decide to remove a few obvious outliers by another threshold on DFO (DFO<60).

A plot of MFL vs. DFO using the second threshold (DFO<60) is shown on Figure 18.

One can see on this plot, as well as on the histogram of MFL and DFO, that a reasonable threshold on MFL is MFL > a, with 2 < a < 5. We decide to test linear regression using both thresholds. The results of these regressions are shown on Table 8.

These results show that a correlation is clear, with a slight underestimation for the model compared to observation. They also confirm, according to the standard error, that no certainty can be obtained in the case of small water height.

With threshold (MFL>5) and (DFO<60), a histogram and a boxplot of the variation (MFL-DFO)/DFO show that a reasonable 95% confidence interval should be [-0.7, 0.7] (even if the non normality of the variable prevents from the standard calculation).

Finally, the results, obtained from this specific case of the Ganges and Brahmaputra basins, show real improvement in term of flood surfaces, particularly in intermediate and large downstream basins. This modified procedure could be included in further efforts to improve this global flood model.

6 Conclusion

This study is a first attempt to use the method of peak flow estimates at a global level. It needed a particular effort in the collection and treatment of data - especially discharge station data - from various sources and in the geomatic processing, due in particular to the use of the very recent 90m HydroSHEDS D.E.M. The statistical analysis showed the possibility of grouping the basins in a global approach. The results of the present study were used, in combination with the DFO dataset of observed flood events, in the Global Assessment Report (Herold and Mouton, 2009; Peduzzi et al., 2010). The combination of these two datasets was relevant for the global risk analysis. However, this final map should neither be applied at a local scale, nor for prediction.

The results of this study allowed identifying different issues and problems, some of which could be partly solved in a further effort for improving the model.

Concerning future improvements, the most important would be to get better, denser and more - temporally and spatially - homogeneous discharge station data. The best would be to use daily time-series in place of monthly ones. These data exist for a lot of countries, but are very uneasy to obtain.

The geomatic treatment could also be improved if the impact of river "burning" on flooded area estimates could be taken into account, which seems not straightforward. Two classes of basins could also be considered, according to their area, in order to get better results on very large ones. The positioning of stations could be realized on the HydroSHEDS 90m D.E.M., in order to improve the quality of statistical variables. This was not possible here since this D.E.M. was not fully available at this stage of the project.

The statistical analysis is widely relying on the quantity and quality of the data. It seems reasonable to think that the improvements in the initial datasets would lead to finer results. Although a more subtle analysis could be done - by considering other distributions, other methods for the estimation of parameters…-, this should be considered – in our opinion – as "fine tuning", with little impact compared to the possible improvements listed above.

7 References

Beck, C., Grieser, J. and Rudolf, B., A New Monthly Precipitation Climatology for the Global Land Areas for the Period 1951 to 2000, Climate Status Report 2004, 181-190, 2005, German Weather Service, Offenbach, Germany.

Bravard, J.-P., Petit, F., Les cours d'eau, Dynamique du système fluvial, 222 p., Armand Colin, Paris, 1997.

Chow, V. T., Maidment, D. R., Mays, L. W., Applied Hydrology, 572 p., McGraw-Hill, New York, 1988.

Danner, A., Mølhave, T., Yi, K., Agarwal, P. K., Arge, L., and Mitasova, H., TerraStream: from elevation data to watershed hierarchies. In Proceedings of the 15th Annual ACM international Symposium on Advances in Geographic information Systems (Seattle, Washington, November 07-09, 2007). GIS '07. ACM, New York, NY, 1-8, http://doi.acm.org/10.1145/1341012.1341049.

Dilley, M., Chen, R., Deichmann, U., Lerner-Lam, A. and Arnold, M., Natural Disaster Hotspot: A Global Risk Analysis, Hazard Management Unit, World Bank, Washington DC, 2005.

Dupuis, D.J., Parameter and quantile estimation for the generalized extreme-value distribution: a second look, Environmetrics, 10, 11-124, 1999.

Farquharson, F.A.K., Meigh, J.R., and Sutcliffe, J.V., Regional flood frequency analysis in arid and semi-arid areas. Journal of Hydrology, 138, 487-501, 1992.

Giuliani, G., Peduzzi, P., The PREVIEW Global Risk Data Platform: A geoportal to serve and share global data on risk to natural hazards, (in prep.).

Herold, C. and Mouton, F., Global Flood Modelling, Statistical Estimation of Peak Flow Magnitude, 41 p., World Bank Development Research Group - UNEP/GRID-Europe, 2006, http://www.grid.unep.ch/product/publication/download/article_global_flood_modeling.pdf

Herold, C. and Mouton, F., Statistical estimates of peak-flow magnitude, contribution in Global Assessment Report on Disaster Risk Reduction. United Nations International Strategy for Disaster Reduction Secretariat, 207 p., 2009, http://www.preventionweb.net/english/hyogo/gar/report/

Kachroo, R.K., Mkhandi, S.H. and Parida, B.P., Flood frequency analysis of southern Africa: I. Delineation of homogeneous regions, Hydrological Sciences, 45 (3), 437-447, 2000.

Katz, R.W., Parlange, M.B., and Naveau, P., Statistics of extremes in hydrology. Advances in Water Resources, 25, 1287-1304, 2002.

Kjeldsen, T.R., Smithers, J.C., and Schulze, R.E., Flood frequency analysis at ungauged sites in the Kwazulu-Natal Province, South Africa. Water SA, 27 (3), 315-324. 2001, http://www.wrc.org.za

Kottek, M., Grieser, J., Beck, C., Rudolf, B. and Rubel, F., World Map of the Köppen-Geiger climate classification updated, Meteorologische Zeitschrift, 15 (3), 259-263, 2006.

Lehner, B. and Döll, P., Development and validation of a global database of lakes, reservoirs and wetlands. Journal of Hydrology, 296 (1-4), 1-22, 2004,

http://www.wwfus.org/science/data/globallakes.cfm

Masutomi, Y., Inui, Y., Takahashi, K. and Matsuoka, Y., Development of highly accurate global polygonal drainage basin data, Hydrological Processes, 23 (4), 572-584, 2009.

McGregor, G.R., Application of regional flood frequency analysis to large tropical catchments: A case study in the Sepik Basin, Papua, New Guinea, Singapore Journal of Tropical Geography, 11 (1), 1- 12,1990.

Meigh, J.R., Farquharson, F.A.K, Sutcliffe, J.V., A worldwide comparison of regional flood estimation methods and climate. Hydrological Sciences, 42 (2), 225-244, 1997.

Mkhandi, S. H., Kachroo, R. K. & Gunasekara, T. A. G. Flood frequency analysis of southern Africa: II. Identification of regional distributions. Hydrol. Sci. J., 45 (3), 449-466, 2000.

Mitchell, T.D. and Jones, P.D., An improved method of constructing a database of monthly climate observation and associated high-resolution GRIDs, Int. J. Climatol., 25, 693–712, 2005.

Musy, A., Hydrologie générale, 2005, http://hydram.epfl.ch/e-drologie/

Musy, A., Soutter, M., Physique du sol, 348 p., Presses polytechniques et universitaires romandes, Lausanne, 1991.

Peduzzi, P., Dao, H., Herold, C., Mouton, F., Assessing global exposure and vulnerability towards natural hazards: the Disaster Risk Index, Nat. Hazards Earth Syst. Sci., 9, 1149–1159, 2009, http://www.nat-hazards-earth-syst-sci.net/9/1149/2009/

Peduzzi, P., Chatenoux, B., Dao, H., De Bono, A., Deichmann, U., Giuliani, G., Herold, C., Kalsnes, B., Kluser, S., Løvholt, F., Lyon, B., Maskrey, A., Mouton, F., Nadim, F. and Smebye, H., The Global Risk Analysis for the 2009 Global Assessment Report on

Disaster Risk Reduction, Extended abstract for International Disaster and Risk Conference, Davos, Switzerland, 2010,

https://www.conftool.com/idrc2010/index.php?page=browseSessions&abstracts=show&form_session=83&presentations=show&downloads=show

Peel, M.C., Finlayson, B.L. and McMahon, T.A., Updated world map of the Köppen-Geiger climate classification, Hydrol. Earth Syst. Sci., 11, 1633–1644, 2007.

Reidy Liermann C., Ecohydrologic impacts of dams : A global assessment. [Thesis]. Umeå: Ekologi, miljö och geovetenskap, 2007.

Sando, S.K., Techniques for Estimating Peak Flow Magnitude and Frequency Relations for South Dakota Streams, Water-Resources Investigation Report 98-4055, USDI/USGS, 1998.

Topaloglu, F., Regional flood frequency analysis of the basins of the East Mediterranean Region, Turk J Agric For, 29, 287-295, 2005.

UNDP/BCPR: A Global Report: Reducing Disaster Risk, A Challenge for Development, New York, 146 p., 2004.

United States Water Resources Council's Hydrology Subcommittee, Bulletin 17B: Guidelines for determining flood flow frequency, 194 p., 1982.

Verdin, K.L., A System for Topologically Coding Global Drainage Basins and Stream Networks, Earth Resources Observation Systems (EROS) Data Center, 1997, http://proceedings.esri.com/library/userconf/proc97/proc97/to350/pap311/p311.htm

Verdin, K.L., Verdin, J.P., A topological system for delineation and codification of Earth's river basins, 218 (1-2), 1-12, 1999.

8 Datasets

8.1 River discharge datasets:

- Long-term mean monthly discharge dataset. The Global Runoff Data Centre (GRDC), 56002 Koblenz, Germany.

  http://grdc.bafg.de/servlet/is/987/

- R-ArcticNET, A Regional, Electronic, Hydrographic Data Network For the Arctic Region. Water Systems Analysis Group. Complex Systems Research Center. Institute for the Study of Earth, Oceans and Space. University of New Hampshire.

  http://www.r-arcticnet.sr.unh.edu/v4.0/index.html

- The Global River Discharge Database (RivDIS v1.1). Water Systems Analysis Group. Complex Systems Research Center. Institute for the Study of Earth, Oceans and Space. University of New Hampshire.

  http://www.rivdis.sr.unh.edu/

- Monthly Discharge Data for World Rivers (except former Soviet Union). DE/FIH/GRDC and UNESCO/IHP, 2001: Monthly Discharge Data for World Rivers (except former Soviet Union). Published by the CISL Data Support Section at the National Center for Atmospheric Research, Boulder, CO (ds552.1). http://dss.ucar.edu/datasets/ds552.1/

- Russian River Flow Data by Bodo, Enhanced. Monthly river flow rates for Russia and former Soviet Union countries in ds553.1 are augmented with data from Russia's State Hydrological Institute (SHI) and a few sites from the Global Hydroclimatic Data Network (GHCDN).

  http://dss.ucar.edu/datasets/ds553.2/

- Discharge of selected rivers of the world. World Water Resources and their use, a joint SHI/UNESCO product. International Hydrological Programme. UNESCO's intergovernmental scientific programme in water resources.

  http://webworld.unesco.org/water/ihp/db/shiklomanov/

- Dados de Base de Rios, Sistema Nacional de Informação de Recursos Hídricos, Instituto da Água, Ministério do Ambiente, do Ordenamento do Território e do Desenvolvimento Regional, Governo da República Portuguesa.

  http://snirh.pt/

- Ecoulements Mensuels Mesurés. Origine des données: AE RMC, CNR, DIREN PACA, DIREN Rhône-Alpes, DIREN Rhône-Alpes + CNR, EDF / HYDRO - MEDD/DE – Données ayant fait l'objet de modifications par un tiers – La responsabilité de la Direction de l'Eau et des producteurs de données ne peut être engagée.

  http://www.hydro.eaufrance.fr/

- Caudal en las Estaciones de Aforo, Confederación Hidrográfica del Duero, Ministerio de Medio Ambiente y Medio Rural y Marino, Gobierno de España.

  http://www.chduero.es/

- Caudal en las Estaciones de Aforo, Confederación Hidrografica del Ebro, Ministerio de Medio Ambiente y Medio Rural y Marino, Gobierno de España.

  http://www.chebro.es/

- Débit quotidien et maximums instantanés annuels, Office fédéral de l'environnement OFEV, Division Hydrologie, 21.3.2008.

  http://www.bafu.admin.ch/org/organisation/00196/index.html?lang=fr

8.2 Digital Elevation Model and hydrological derived datasets:

- HYDRO1k Elevation Derivative Database. EROS, USGS. http://edc.usgs.gov/products/elevation/gtopo30/hydro/index.html
- HydroSHEDS, WWF. In partnership with USGS, CIAT, TNC, CESR. http://www.worldwildlife.org/hydrosheds
- Global Drainage Basin Database (GDBD). Yuji Masutomi, Yusuke Inui, Kiyoshi Takahashi, and Yuzuru Matsuoka (2007) Development of highly accurate global polygonal drainage basin data. Submitted to Hydrological Processes.

  http://www-cger.nies.go.jp/cger-e/db/enterprise/gdbd/gdbd_index_e.html

8.3 Land cover datasets:

- Global land cover GLC_2000 version 1. Institute for Environment and Sustainability, Joint Research Centre.

  http://www-gvm.jrc.it/glc2000/

- Global Lakes and Wetlands Database (GLWD). Lehner, B. and Döll, P. (2004): Development and validation of a global database of lakes, reservoirs and wetlands. Journal of Hydrology 296/1-4: 1-22. http://www.wwfus.org/science/data/globallakes.cfm

8.4 Climatic datasets:

- CRU TS 2.1 monthly precipitation. Mitchell, T.D., 2004: An improved method of constructing a database of monthly climate observations and associated high-resolution grids.

  http://www.cru.uea.ac.uk/~timm/grid/CRU_TS_2_1.html

- CRU TS 2.1 monthly mean temperatures. Mitchell, T.D., 2004: An improved method of constructing a database of monthly climate observations and associated high-resolution grids.

  http://www.cru.uea.ac.uk/~timm/grid/CRU_TS_2_1.html

- Variability Analyses of Surface Climate Observations (VASClimO) at the Global Precipitation Climatology Centre (GPCC). Version-1.1, 0.5°x0.5°. Beck,C., Grieser, J. and Rudolf B. (2005): A New Monthly Precipitation Climatology for the Global Land Areas for the Period 1951 to 2000, Climate Status Report 2004, pp. 181 - 190, German Weather Service, Offenbach, Germany.

  http://www.dwd.de

- World Map of the Köppen-Geiger climate classification updated. Kottek, M., J. Grieser, C. Beck, B. Rudolf, and F. Rubel, 2006: World Map of the Köppen-Geiger climate classification updated. *Meteorol. Z.*, **15**, 259-263. University of Veterinary Medicine Vienna.

  DOI: 10.1127/0941-2948/2006/0130

- Updated world map of the Köppen-Geiger climate classification. Peel MC, Finlayson BL & McMahon TA (2007), Updated world map of the Köppen-Geiger climate classification, Hydrol. Earth Syst. Sci., 11, 1633-1644. The University of Melbourne, Victoria, Australia.

  http://www.hydrol-earth-syst-sci.net/11/1633/2007/hess-11-1633-2007.html

- The Holdridge Life Zones data set. Leemans, Rik, 1990. Global data sets collected and compiled by the Biosphere Project, Working Paper, IIASA-Laxenburg, Austria.

  http://www.grid.unep.ch/data/data.php?category=biosphere

8.5 Recorded flood event dataset:

- World Atlas of Flooded Lands. Dr. G. Robert Brakenridge, Ms. Elaine Anderson. Dartmouth Flood Observatory.

http://www.dartmouth.edu/~floods/index.html

9 Tables

Table 1 Independent variables generated for regression analysis.

	Variable	Description	Abbreviation
	Hydromorphometric		
1	Drainage area	Area of drainage basin (km^2).	DRAREA
2	Mean basin elevation	Mean elevation of drainage basin (m).	MEANALT
3	Mean basin slope	Mean slope of drainage basin (m/km).	MNSLOP
4	Basin shape	Gravelius coefficient of compacity (Kc): ratio of basin perimeter to the circle of equal area.	KGRAV
5	Main channel length	Total length of basin main channel (km).	MCHLENGTH
6	Main channel slope	Maximum difference in elevation of basin main channel divided by channel length (m/km).	MCHSLOPE
7	Drainage frequency	Number of Strahler first order streams per square km in basin ($1/km^2$).	DRFREQ
	Land cover		
8	Surface water storage	Cumulated area of every lake and reservoir contained in GLWD level 3. Variable expressed as a ratio to the basin drainage area.	WATER_STOR
9	Forest cover	Global land cover GLC_2000 version 1: cumulated area of any "Tree Cover" classes and the class "Tree Cover / Other natural vegetation". Variable expressed as a ratio to the basin drainage area.	TFORCOV
10	Impervious cover	Global land cover GLC_2000 version 1: area of class 22 “Artificial surfaces and associated areas”. Expressed as a ratio to the basin drainage area.	URBCOV
	Climatic time-series		
11	Mean annual precipitation	Calculated using CRU TS 2.1 dataset on the 1953-2002 period (mm).	PRMEAN
12	Minimum mean monthly temperature	Calculated using CRU TS 2.1 dataset on the 1953-2002 period (°C).	CLDERMONTH
13	Monthly maximum precipitation for a 100-year return period.	Log-Pearson type III estimates using Variability Analyses of Surface Climate Observations (VASClimO) at the Global Precipitation Climatology Centre (GPCC). Version-1.1, 0.5°x0.5°, (mm).	LogP100
	Climatic zones		
14	Percentage area of Köppen-Geiger climatic zones.	Calculated using the World Map of the Köppen-Geiger climate classification updated. University of Veterinary Medicine, Vienna (Kottek et al., 2006).	Koge5
15	Percentage area of Köppen-Geiger climatic zones.	Calculated using the Updated world map of the Köppen-Geiger climate classification. The University of Melbourne, Victoria, Australia (Peel et al., 2006).	Koge1
16	Percentage area of Holdridge climatic zones.	Calculated using the Holdridge Life Zones. IIASA-Laxenburg, Austria.	Holdridge

Table 2 Description of Köppen-Geiger climate symbols and defining criteria.

1st	2nd	3rd	Description	Criteria*
A			Tropical	Tcold≥18
	f		-Rainforest	Pdry≥60
	m		-Monsoon	No t(Af) & Pdry≥100–MAP/25
	w		-Savannah	Not (Af) & Pdry<100–MAP/25
B			Arid	MAP<10×Pthreshold
	W		-Desert	MAP<5×Pthreshold
	S		-Steppe	MAP≥5×Pthreshold
		h	-Hot	MAT≥18
		k	-Cold	MAT<18
C			Temperate	Thot>10 & 0<Tcold<18
	s		-Dry Summer	Psdry<40 & Psdry<Pwwet/3
	w		-Dry Winter	Pwdry<Pswet/10
	f		-Without dry season	Not (Cs) or (Cw)
		a	-Hot Summer	Thot≥22
		b	-Warm Summer	Not (a) & Tmon10≥4
		c	-Cold Summer	Not (a or b) & 1≤Tmon10<4
D			Cold	Thot>10 & Tcold≤0
	s		-Dry Summer	Psdry<40 & Psdry<Pwwet/3
	w		-Dry Winter	Pwdry<Pswet/10
	f		-Without dry season	Not (Ds) or (Dw)
		a	-Hot Summer	Thot≥22
		b	-Warm Summer	Not (a) & Tmon10≥4
		c	-Cold Summer	Not (a,b or d)
		d	-Very Cold Winter	Not (a or b) & Tcold<–38
E			Polar	Thot<10
	T		-Tundra	Thot>0
	F		-Frost	Thot≤0

*MAP = mean annual precipitation, MAT = mean annual temperature, Thot = temperature of the hottest month, Tcold = temperature of the coldest month, Tmon10 = number of months where the temperature is above 10, Pdry = precipitation of the driest month, Psdry = precipitation of the driest month in summer, Pwdry = precipitation of the driest month in winter, Pswet = precipitation of the wettest month in summer, Pwwet = precipitation of the wettest month in winter, Pthreshold = varies according to the following rules (if 70% of MAP occurs in winter then Pthreshold = 2 x MAT, if 70% of MAP occurs in summer then Pthreshold = 2 x MAT + 28, otherwise Pthreshold = 2 x MAT + 14). Summer (winter) is defined as the warmer (cooler) six month period of ONDJFM and AMJJAS.

Table 3 Description of the 22 final groups

1	Af
2	Am
3	Aw, South America
4	Aw, Africa
5	Aw, North America, Europe, Asia and Australia
6	BWh
7	BWk
8	BSh
9	BSk
10	Csa
11	Csb and Csc
12	Cwa, Cwb and Cwc
13	Cfa, Cfb and Cfc
14	Dsa , Dsb, Dsc and Dsd
15	Dwa, Dwb, Dwc and Dwd
16	Dfa and Dfb, North America
17	Dfa and Dfb, Europe
18	Dfa and Dfb, Asia
19	Dfc and Dfd, North America
20	Dfc and Dfd, Europe
21	Dfc and Dfd, Asia
22	ET and EF

Table 4 Summary of regressions estimating peak flows given basin variables

Group	**1**	**2**	**3**	**4**	**5**	**7**	**8**	**9**	**10**	**11**	**12**
N	24	41	133	70	42	17	63	68	23	36	90
Constant	-1.0136	-4.3814	-2.9401	-5.732	-3.8151	-3.3024	-5.7843	-4.9801	-1.7871	-4.767	-4.8795
(p-value)	(0.066)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.016)	(0.000)	(0.000)
LDRAREA	0.9738	0.92146	0.88171	0.80876	0.93366	1.3954	0.7556	0.8646	0.7937	0.9128	0.91392
(p-value)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
LMEANALT											
(p-value)											
LMNSLOPE	0.2988							0.6767	0.8713		0.34084
(p-value)	(0.018)							(0.000)	(0.000)		(0.000)
LPRMEAN			0.6964	1.6794						1.3915	
(p-value)			(0.001)	(0.000)						(0.000)	
CLDMONTH											
(p-value)											
LOGP100		1.3335			1.1556		2.0401	1.2361			1.3633
(p-value)		(0.000)			(0.001)		(0.000)	(0.000)			(0.000)
S	0.32315	0.20552	0.32523	0.23722	0.22909	0.21756	0.40775	0.37549	0.30396	0.32999	0.29031
R^2	76.9%	82.6%	62.1%	61.2%	73.6%	92.0%	56.8%	65.4%	73.1%	65.8%	78.4%
R^2-adj	74.7%	81.7%	61.5%	60.1%	72.3%	91.4%	55.3%	63.7%	70.4%	63.7%	77.7%

Group	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
N	218	39	93	193	111	63	144	85	128	19
Constant	-5.1112	-9.750	-6.4232	-5.3693	-5.7653	-3.0864	-1.7614	-1.7024	-1.4298	-0.9983
(p-value)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
LDRAREA	0.93801	1.0785	0.99682	0.90200	0.98955	0.7955	1.06787	0.70636	0.99447	0.98377
(p-value)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
LMEANALT								0.45849		
(p-value)								(0.000)		
LMNSLOPE						0.29007	0.19744		0.21945	
(p-value)						(0.000)	(0.000)		(0.000)	
LPRMEAN	1.38156	2.7940	1.7509	1.5422	1.4294	0.8573				
(p-value)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)				
CLDMONTH		-0.02618	-0.01460		-0.03008			-0.02651		
(p-value)		(0.000)	(0.000)		(0.000)			(0.000)		
LOGP100										
(p-value)										
S	0.21461	0.23755	0.21398	0.27679	0.16717	0.29124	0.28611	0.17898	0.20206	0.10074
R^2	78.0%	77.3%	84.8%	62.0%	87.1%	63.5%	66.5%	74.2%	80.7%	94.9%
R^2-adj	77.8%	75.3%	84.3%	61.6%	86.7%	61.6%	66.0%	73.2%	80.3%	94.6%

Table 5 Independent variables selected by the regressions.

	Variable	**Abbreviation**
1	Drainage area	DRAREA
2	Mean basin elevation	MEANALT
3	Mean basin slope	MNSLOP
4	Mean annual precipitation	PRMEAN
5	Minimum mean monthly temperature	CLDERMONTH
6	Monthly maximum precipitation for a 100-year return period	LogP100
7	Percentage area of Köppen-Geiger climatic zones	Koge1

Table 6 Distribution of gauging stations by continent along treatment.

Region	Available stations	Moved stations	Final subset
Africa	468	457 (97.6%)	136 (29.1%)
Asia	1546	1454 (94.0%)	397 (25.7%)
Australia	283	271 (95.8%)	77 (27.2%)
Europe	1984	1883 (94.9%)	293 (14.8%)
North America	1760	1654 (94.0%)	530 (30.1%)
South America	1511	1421 (94.0%)	311 (20.6%)
Total	**7552**	**7140 (94.5%)**	**1744 (23.1%)**

Available stations: available station with 1000 km^2 and 7 year of record, duplicate between used datasets are removed.

Moved stations: station adjusted on HYDRO1k stream network using drainage area.

Final subset: subset for statistical analysis.

Table 7 Zones showing no or doubtful results, distribution by country.

First 40 countries ordered by decreasing total percentage. Listed countries are larger than 1000 km^2, and surface (km^2) of cumulated four zones is between 5 and 95 % of country total surface (A=Doubtful results (Arid Desert Cold); B=Outside HydroSHEDS coverage; C=Basin<1000 km^2; D= No regression (Arid Desert Hot)).

COUNTRY	A	%	B	%	C	%	D	%	TOTAL %
Bahamas	0	0.0	435	3.1	12,815	90.9	0	0.0	94
Algeria	52,727	2.3	480	0.0	310,930	13.4	1,738,448	75.0	90.8
Niger	0	0.0	0	0.0	74,161	6.3	996,270	84.2	90.4
Tunisia	0	0.0	252	0.2	37,775	24.3	100,398	64.7	89.2
Sweden	0	0.0	372,131	82.8	21,568	4.8	0	0.0	87.6
Cyprus	0	0.0	59	0.7	7,242	80.4	0	0.0	81
Timor-Leste	0	0.0	158	1.1	11,336	77.1	0	0.0	78.1
Jordan	29,006	32.6	3	0.0	3,742	4.2	36,401	40.9	77.6
Haiti	0	0.0	75	0.3	20,684	76.4	0	0.0	76.7
Somalia	0	0.0	421	0.1	62,257	9.8	397,071	62.7	72.6
Iraq	0	0.0	9	0.0	20,133	4.6	291,138	66.8	71.4
Eritrea	0	0.0	75	0.1	30,670	25.5	54,389	45.2	70.7
Turkmenistan	243,639	51.6	797	0.2	80,771	17.1	0	0.0	68.9
Denmark	0	0.0	1,341	3.0	28,791	64.8	0	0.0	67.8
Mali	0	0.0	0	0.0	130,971	10.5	703,276	56.2	66.6
Chad	0	0.0	0	0.0	59,567	4.7	779,124	61.3	66
Pakistan	115,286	13.2	967	0.1	26,189	3.0	409,596	46.8	63
Chile	214,651	28.5	5,841	0.8	249,705	33.1	0	0.0	62.3
Cuba	0	0.0	762	0.7	68,114	61.1	0	0.0	61.8
Afghanistan	215,006	33.5	332	0.1	18,678	2.9	161,011	25.1	61.6
Uzbekistan	230,900	51.4	147	0.0	45,978	10.2	0	0.0	61.6
Philippines	0	0.0	1,108	0.4	176,168	59.5	0	0.0	59.9
Morocco	9,474	2.3	710	0.2	33,087	8.1	198,345	48.8	59.5
Israel	0	0.0	1	0.0	4,747	22.9	7,082	34.1	57
Russian Federation	14,338	0.1	9,355,666	55.2	255,185	1.5	0	0.0	56.8
Namibia	9,729	1.2	530	0.1	61,201	7.4	383,387	46.5	55.2
Australia	163,841	2.1	1,559	0.0	1,087,947	14.2	2,940,301	38.2	54.5
Sudan	0	0.0	96	0.0	91,578	3.7	1,226,788	49.3	53
Greece	0	0.0	297	0.2	65,971	49.8	0	0.0	50
Lebanon	0	0.0	0	0.0	4,944	48.8	0	0.0	48.8
Canada	0	0.0	4,338,059	43.9	462,390	4.7	0	0.0	48.6
Panama	0	0.0	257	0.3	35,372	47.1	0	0.0	47.5
Dominican Republic	0	0.0	93	0.2	22,237	46.2	0	0.0	46.3
Taiwan	0	0.0	232	0.6	16,030	44.4	0	0.0	45
Belize	0	0.0	367	1.6	9,190	41.3	0	0.0	42.9
Japan	0	0.0	2,055	0.5	148,381	39.7	0	0.0	40.3
Syrian Arab Republic	8,495	4.5	11	0.0	11,494	6.1	55,792	29.6	40.3
Tajikistan	53,232	37.5	0	0.0	120	0.1	0	0.0	37.5
Iran	92,153	5.7	453	0.0	108,923	6.7	368,590	22.7	35.2
Mongolia	323,112	20.6	0	0.0	226,826	14.5	0	0.0	35.1

Table 8 Modelized vs. observed average flood height in Ganges basin: results of linear regression.

	R^2-adj	Std error	Coef	Coef Std error	p-value
MFL > 2	0.87	5.7	0.92	0.013	0.000
MFL > 5	0.80	6.6	0.83	0.014	0.000

10 Figures

Figure 1 Global flow chart.

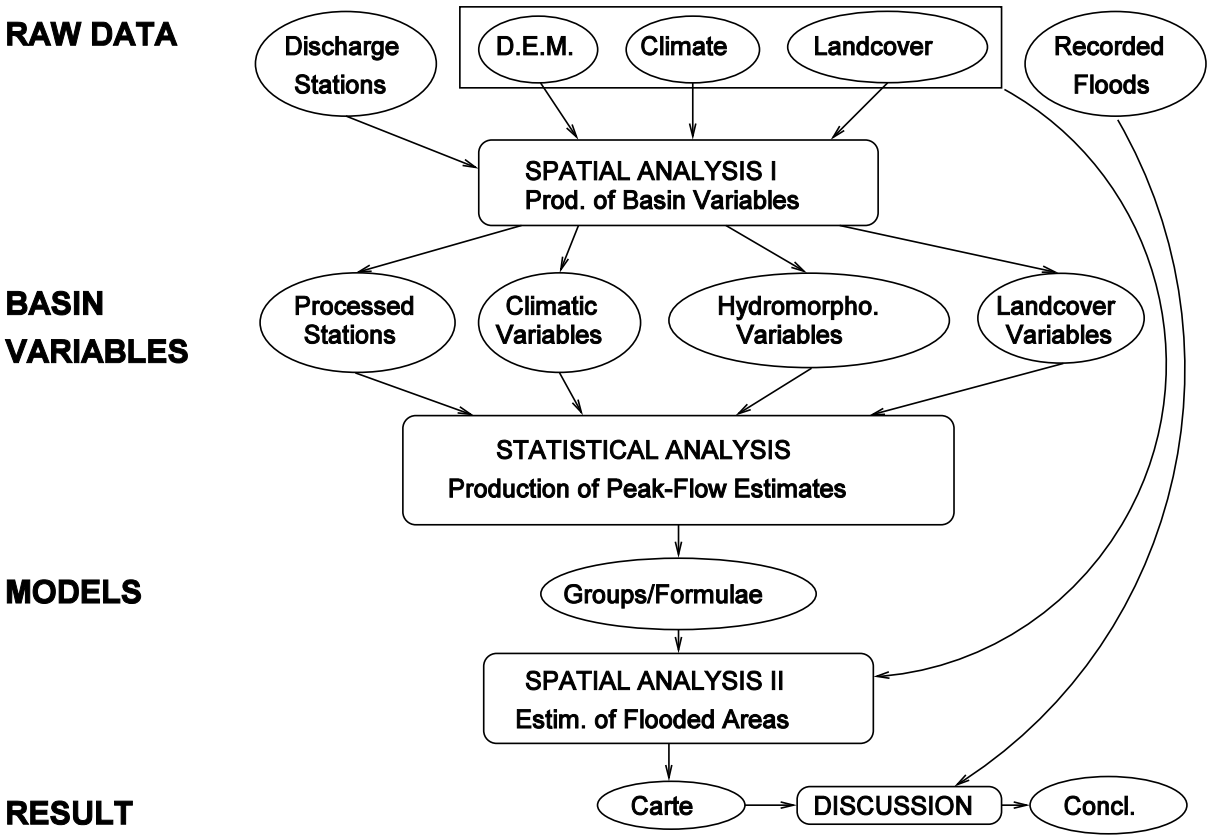

Figure 2 Spatial Analysis I: Production of basin variables.

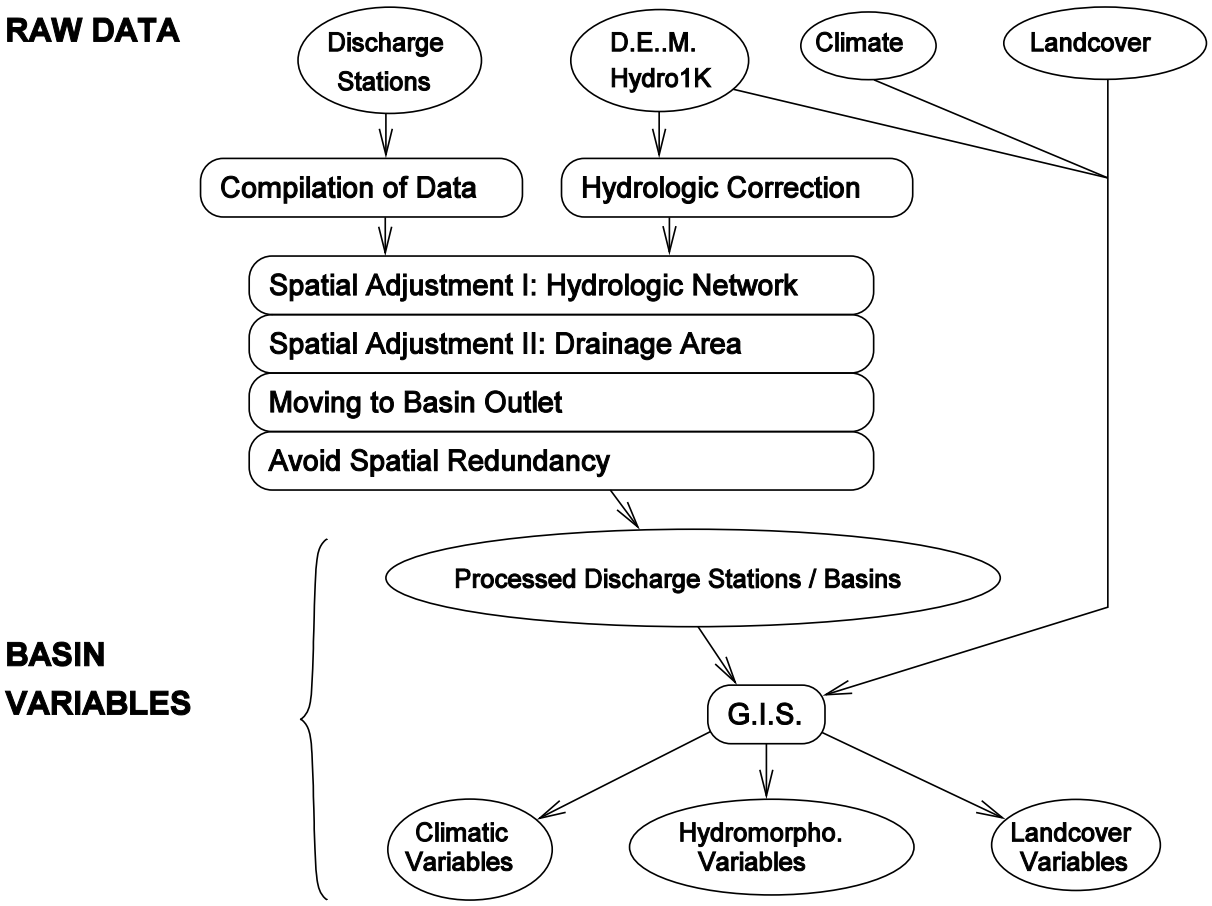

Figure 3 Elaboration of Peak Flow statistical models.

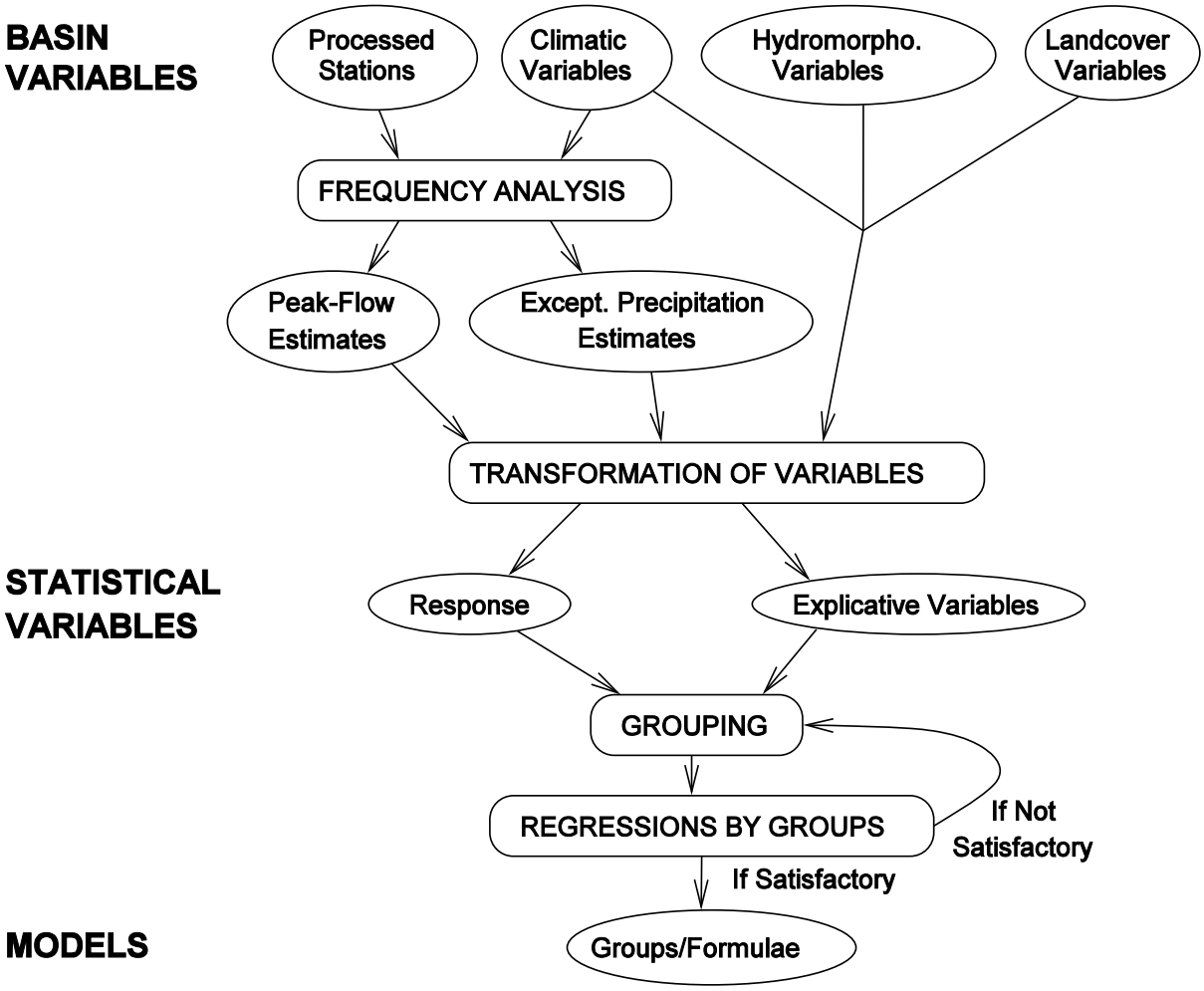

Figure 4 Spatial Analysis II: Estimation of flooded areas.

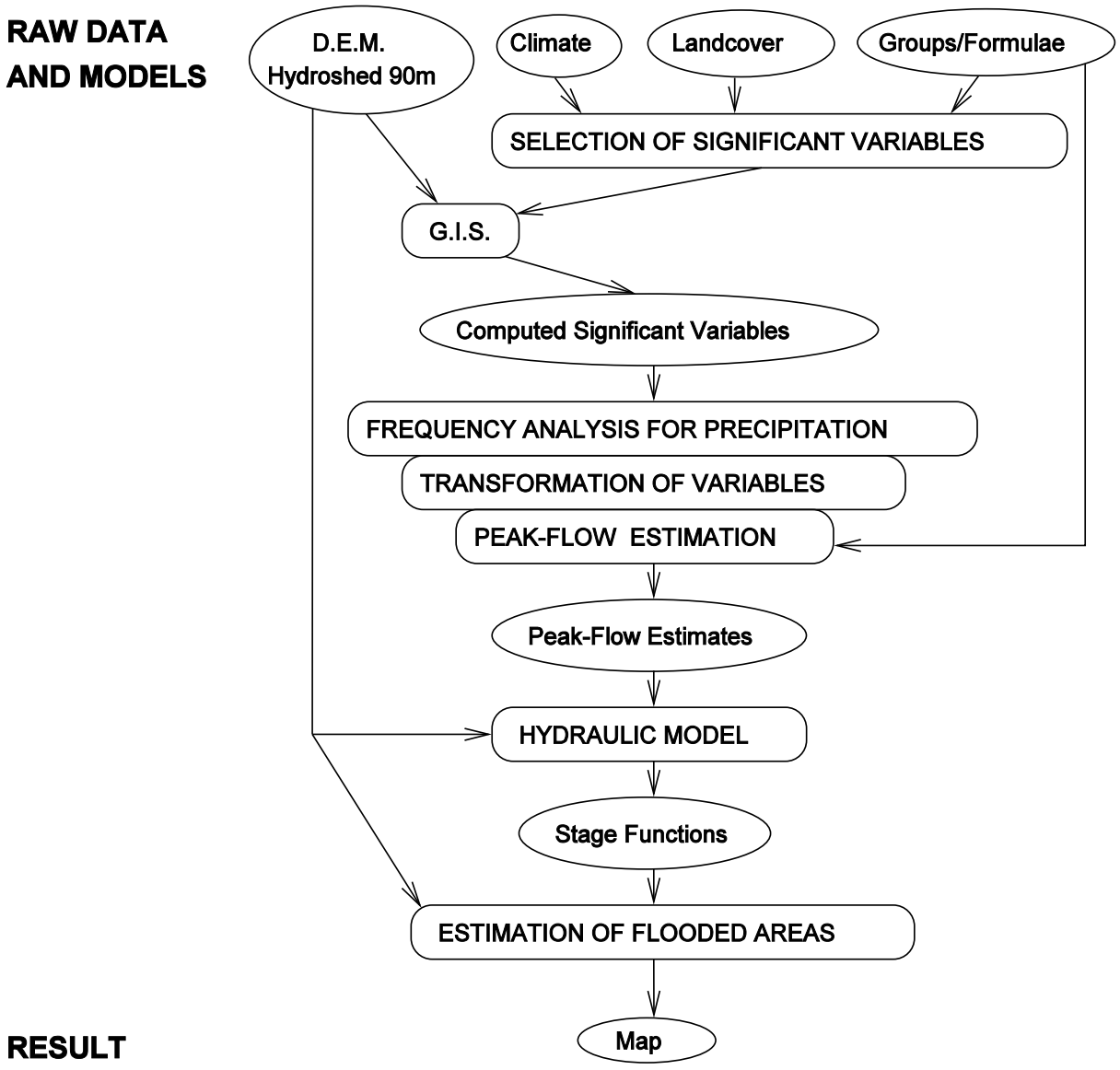

Figure 5 Global distribution of available gauging stations.

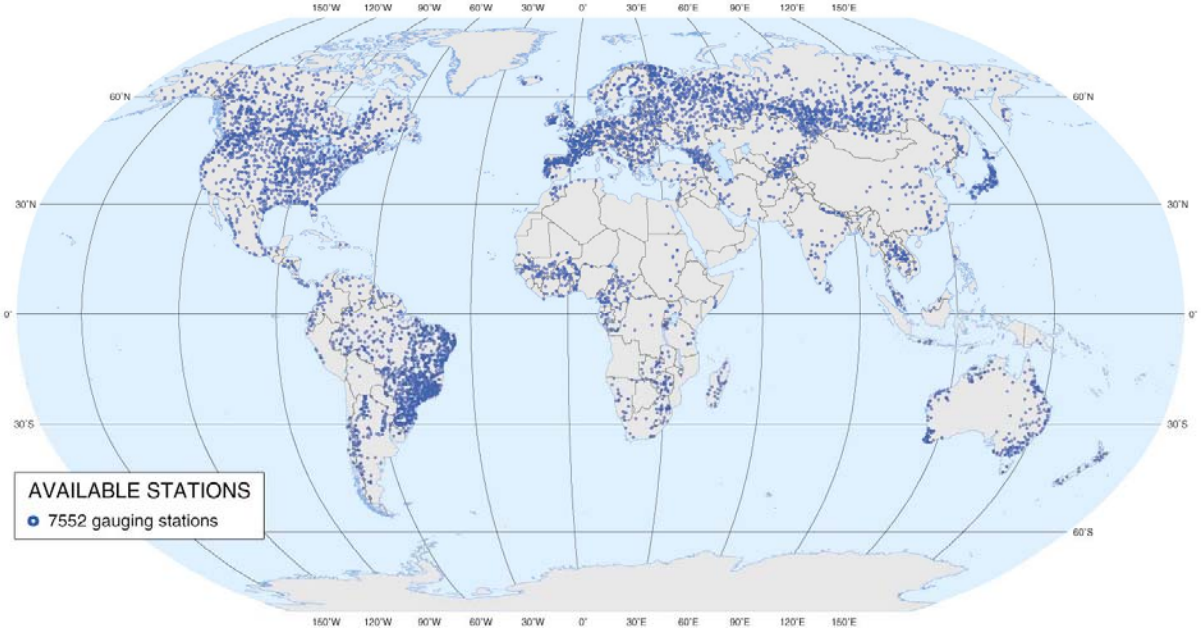

Figure 6 Distribution of available gauging stations by continent.

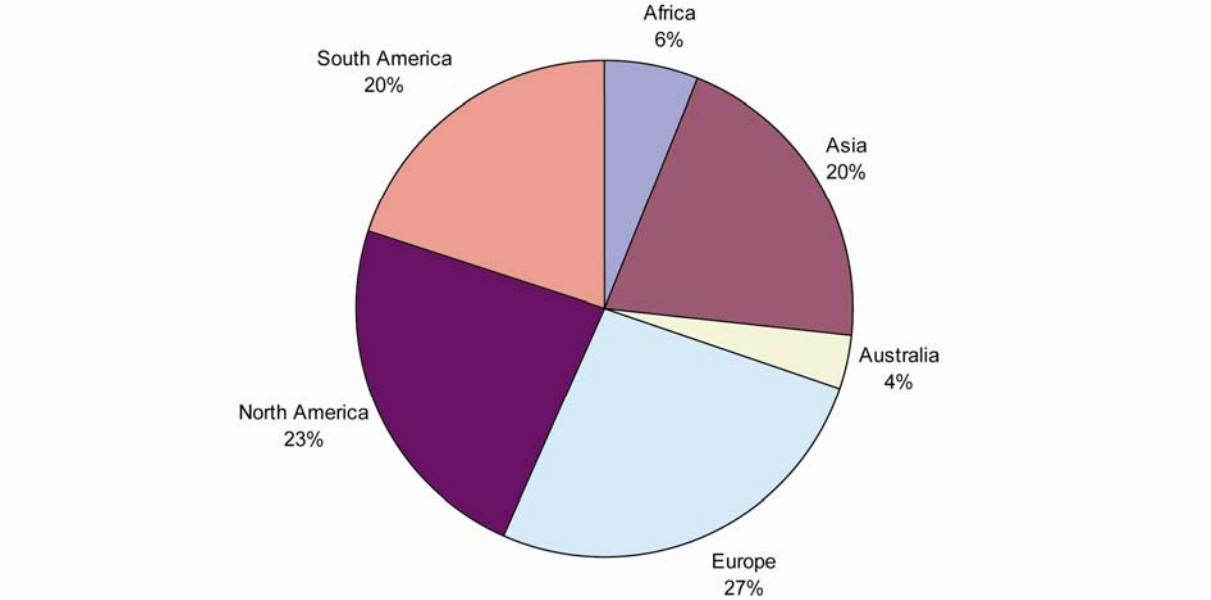

Figure 7 Global distributions of stations final subset and corresponding drainage basins.

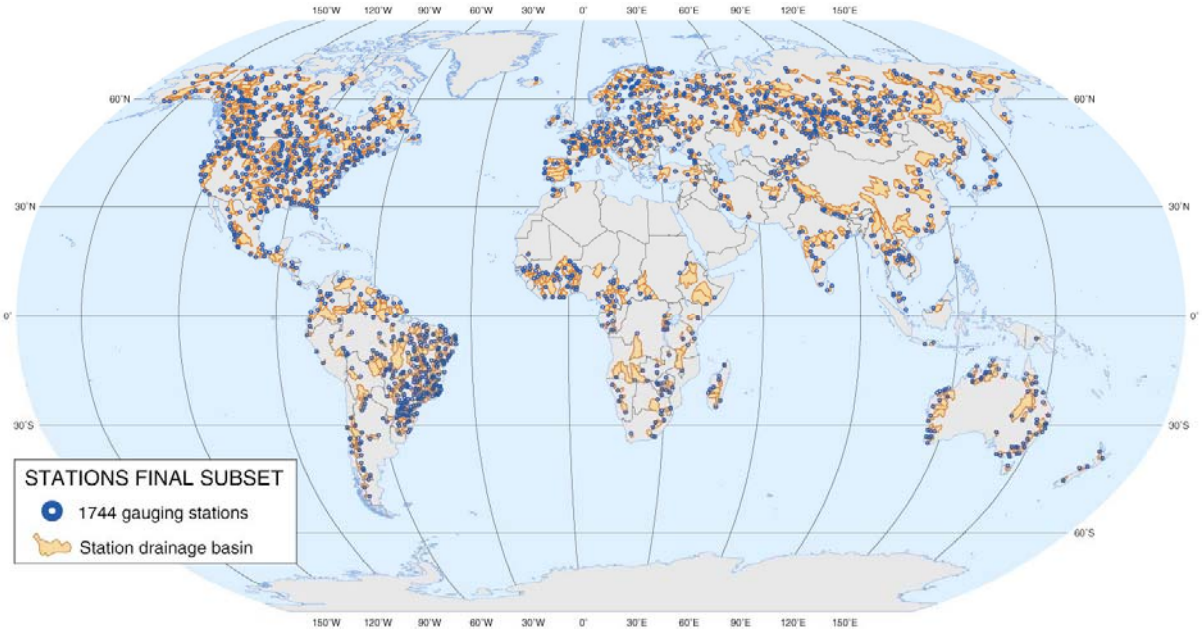

Figure 8 Distribution of gauging stations final subset by continent.

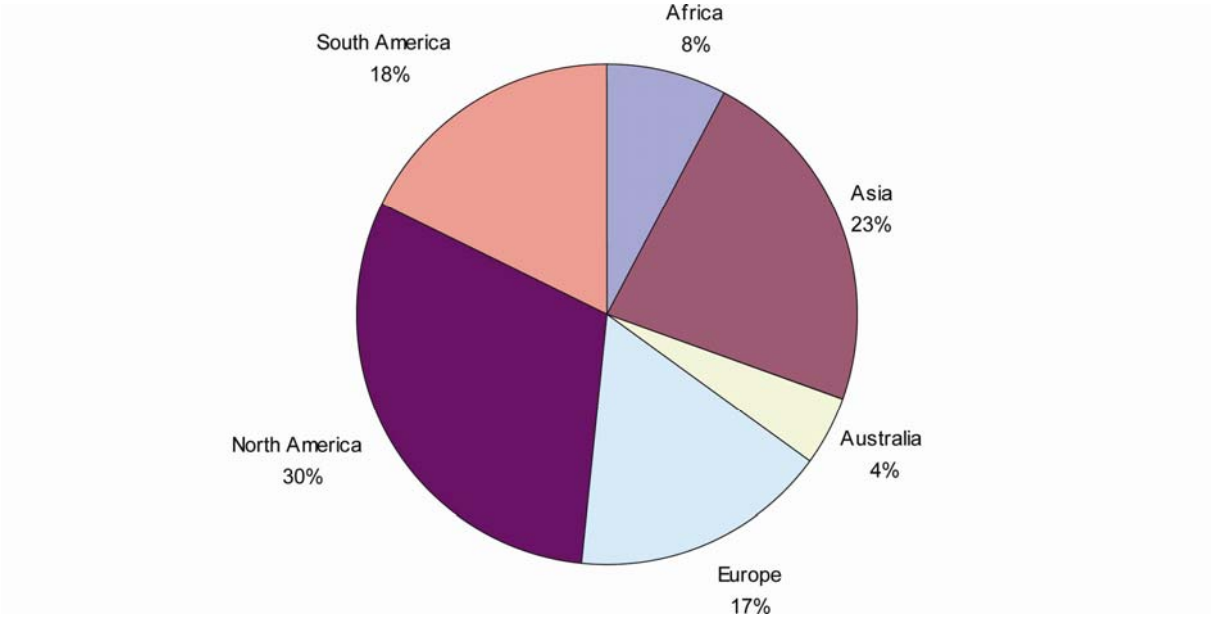

Figure 9 Q100 [m^3/s]: Estimated vs Stations for Ganges basin.

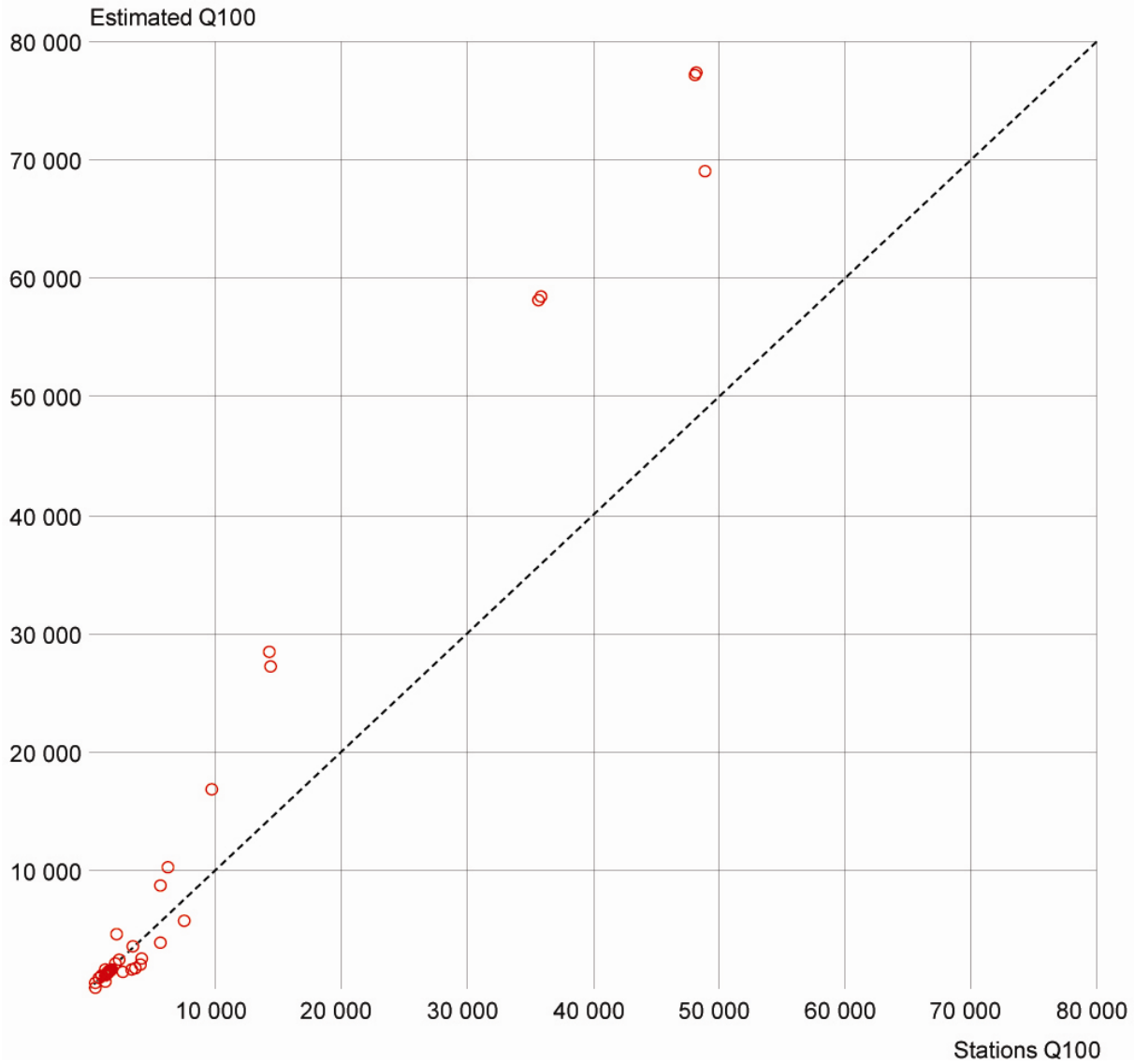

Figure 10 Q100 [m^3/s]: Estimated vs Stations for Yangtze basin.

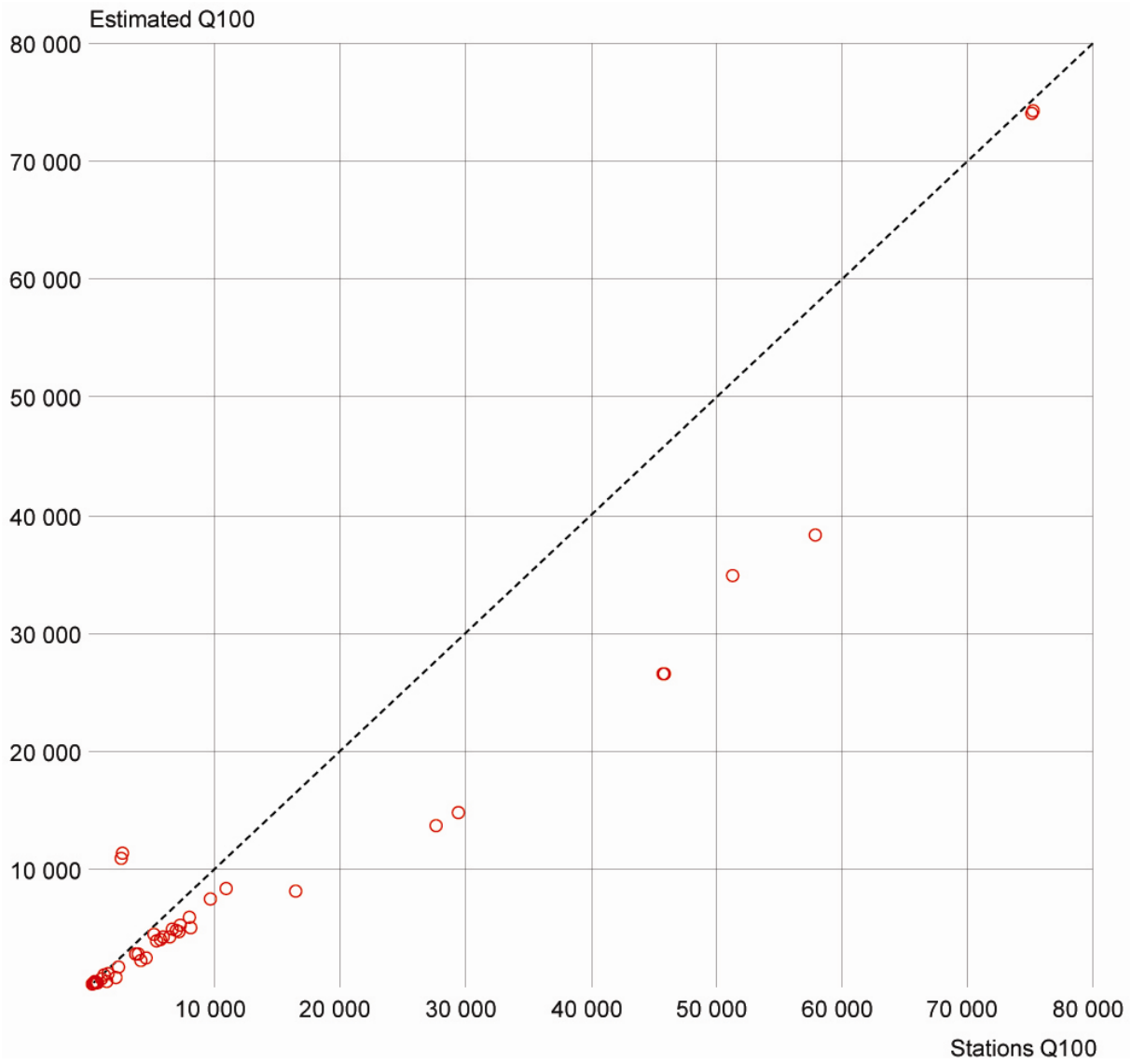

Figure 11 Global distribution of zones showing results, doubtful results or no data.

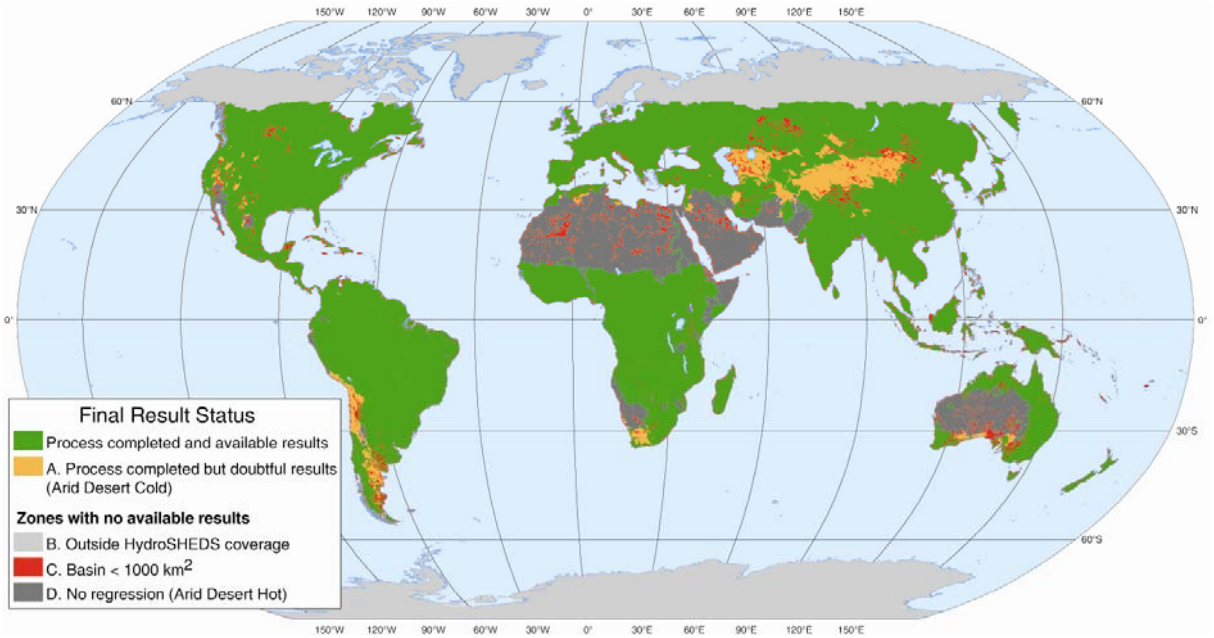

Figure 12 Krishna and Godavari's mouth area.

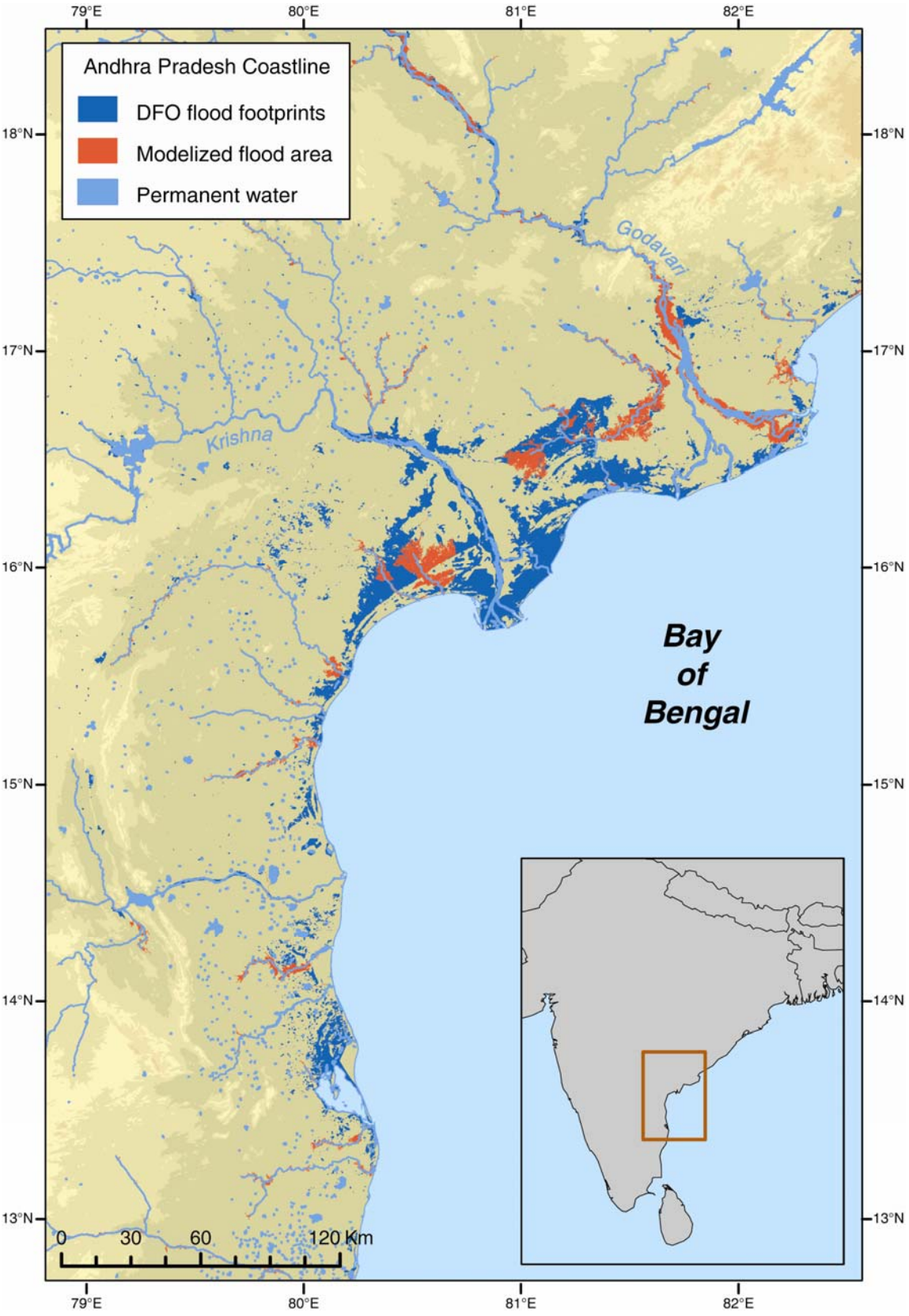

Figure 13 Event cumulated surface and modelized flooded area in the Ganges basin.

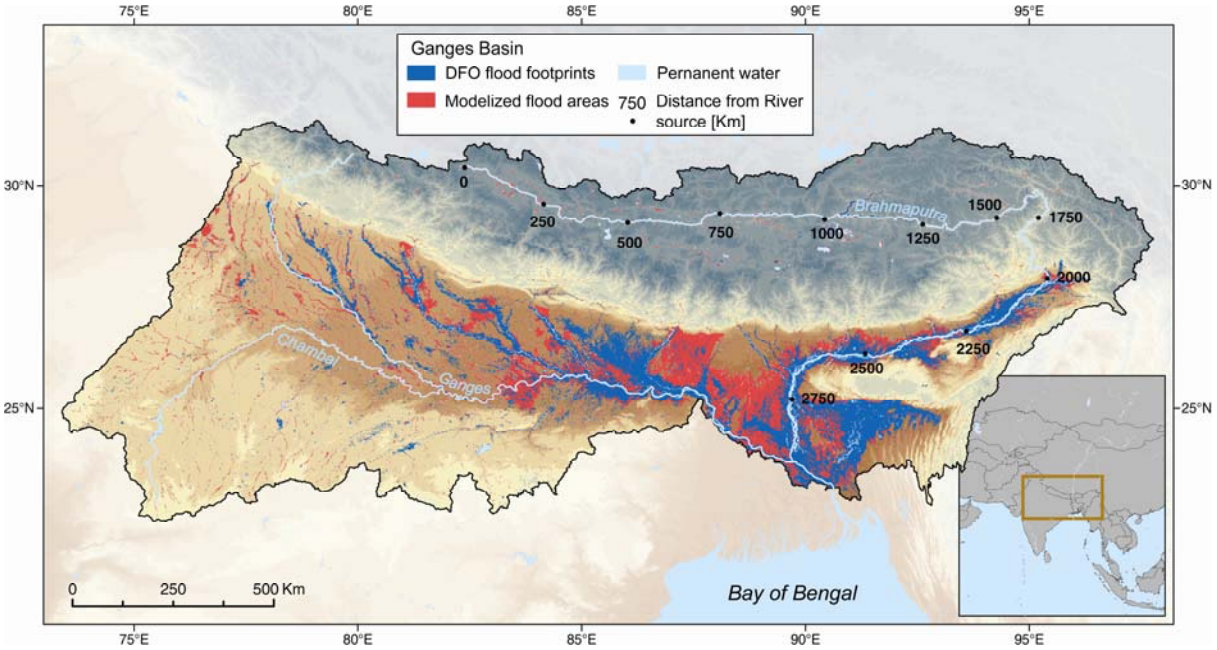

Figure 14 Maximum extent of flooded pattern along Brahmaputra River.

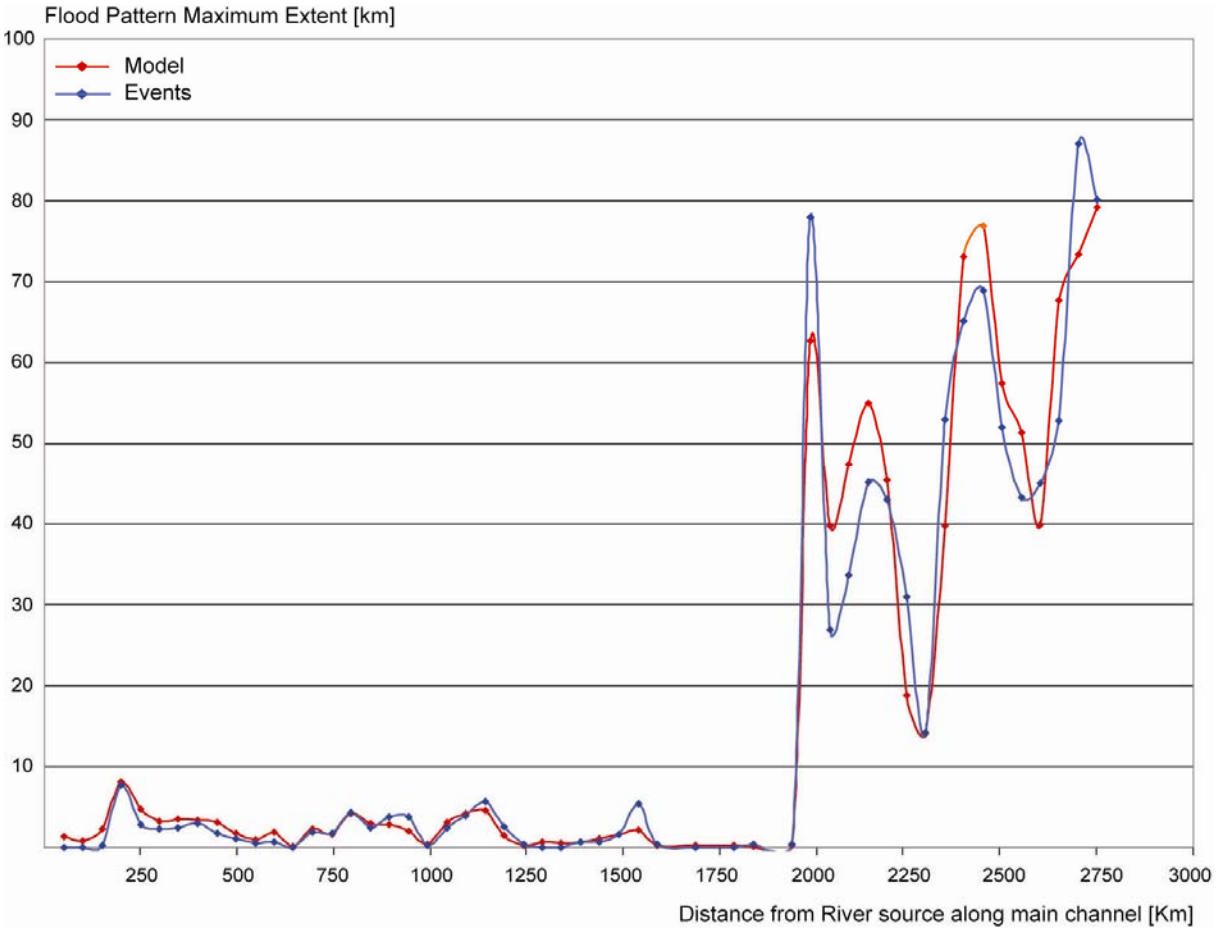

Figure 15 Total surface of flooded area along Brahmaputra River.

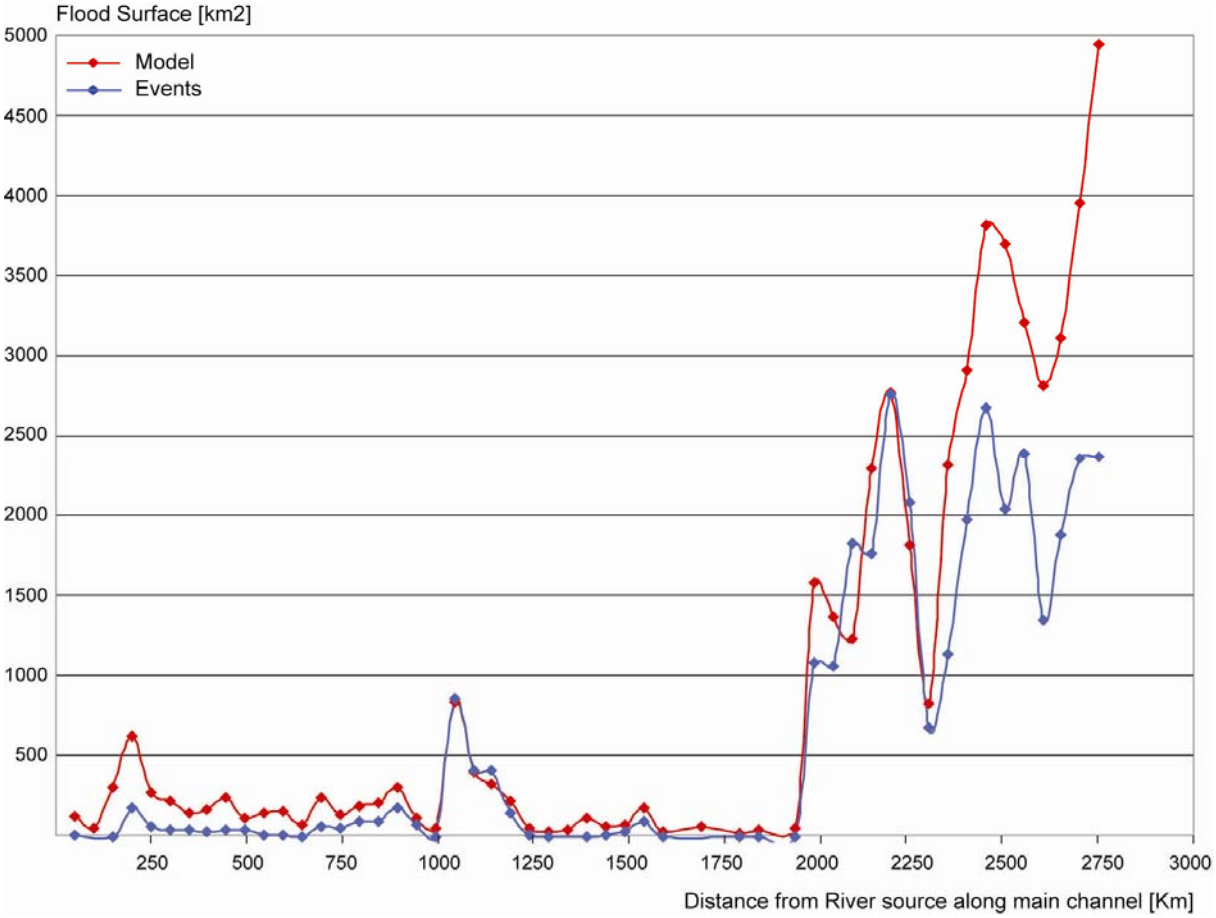

Figure 16 Ganges basin: Percentage of flood event total area not covered by the model, for each Pfafstetter level 4 basin.

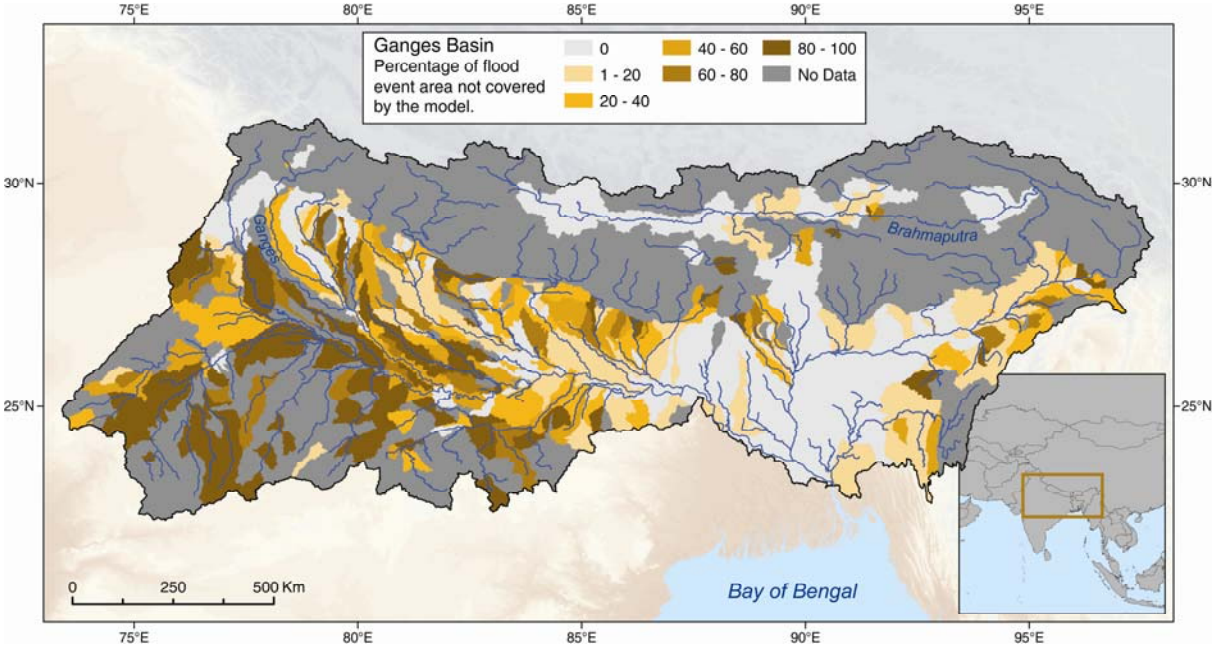

Figure 17 Ganges basin: Ratio of Model to Event average flood height for each river section basin.

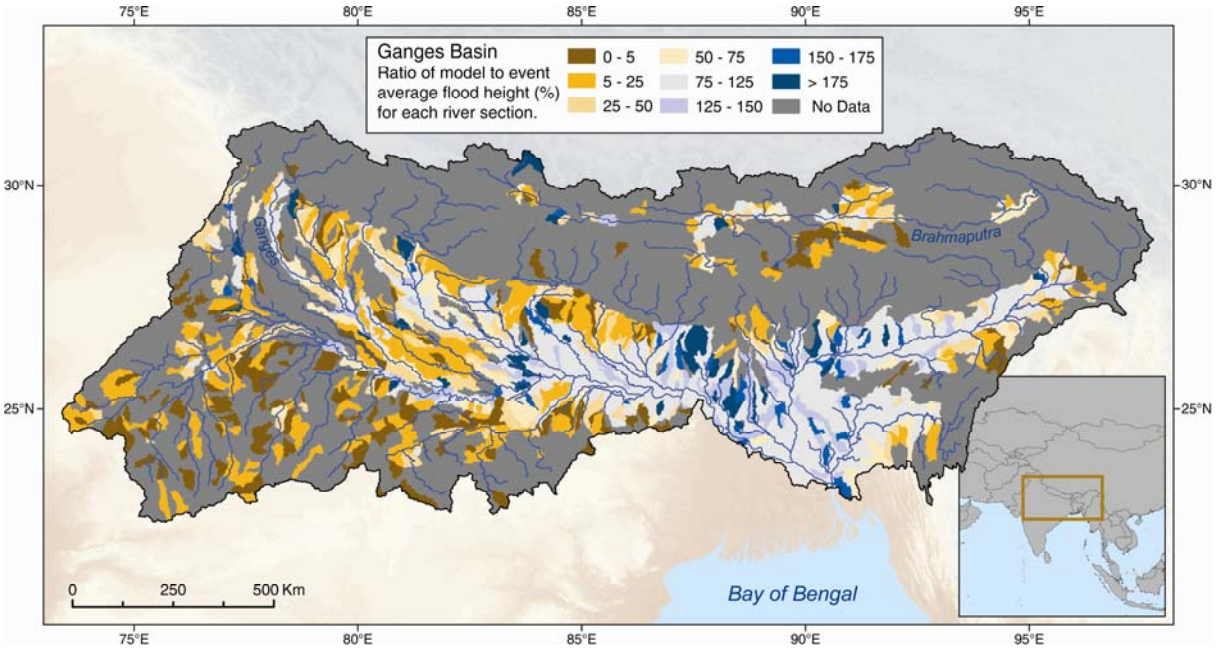

Figure 18 Ganges basin: Model vs Events average flood height [m] for each stream section basin.