Evaluation of the satellite-based Global Flood Detection System for measuring river discharge: Influence of local factors

B. Revilla-Romero [1,2], J. Thielen[1], P. Salamon[1], T. De Groeve[1], G. R. Brakenridge[3]

[1] European Commission Joint Research Centre, Ispra, Italy

[2] Utrecht University, Faculty of Geosciences, Utrecht, the Netherlands

[3] University of Colorado, Boulder, USA

Correspondence to: B. Revilla-Romero (beatriz.revilla-romero@jrc.ec.europa.eu)

Abstract

One of the main challenges for global hydrological modelling is the limited availability of observational data for calibration and model verification. This is particularly the case for real time applications. This problem could potentially be overcome if discharge measurements based on satellite data were sufficiently accurate to substitute for ground-based measurements. The aim of this study is to test the potentials and constraints of the remote sensing signal of the Global Flood Detection System for converting the flood detection signal into river discharge values.

The study uses data for 322 river measurement locations in Africa, Asia, Europe, North America and South America. Satellite discharge measurements were calibrated for these sites and a validation analysis with in situ discharge was performed. The locations with very good performance will be used in a future project where satellite discharge measurements are obtained on a daily basis to fill the gaps where real time ground observations are not available. These include several international river locations in Africa: Niger, Volta and Zambezi rivers.

Analysis of the potential factors affecting the satellite signal was based on a classification decision tree (Random Forest) and showed that mean discharge, climatic region, land cover and upstream catchment area are the dominant variables which determine good or poor performance of the measurement sites. In general terms, higher skill scores were obtained for locations with one or more of the following characteristics: a river width higher than 1km, a large floodplain area and in flooded forest, with a potential flooded area greater than 40%, sparse vegetation, croplands or grasslands and closed to open and open forest, Leaf Area Index > 2, tropical climatic area, and without hydraulic infrastructures. Also, locations where river ice cover is seasonally present obtained higher skill scores. The work provides guidance on the best locations and limitations for estimating discharge values from these daily satellite signals.

Keywords: Floods; Passive microwave sensing; Discharge measurement; global evaluation; local

1 Introduction

Flooding is the most prevalent natural hazard at the global scale, often with dire humanitarian and economic effects. According to the International Disaster Database (EM-DAT), an average of 175 flood events per year occurred globally between 2002-2011, affecting an average of 116.5 million people, and causing economic losses of US$25.5 billion. According to MunichRe (2014), the costliest natural catastrophe worldwide in terms of overall economic losses in 2013 was the flooding in southern and eastern Germany and neighbouring states in May and June with estimated damages of $15.2 billion. In June of the same year, flooding in India cost 5000 lives, with a further 2 million affected (MunichRe, 2014; EM-DAT).

The Global Assessment Report (UNISDR, 2011) states that the proportion of world population living in flood-prone river basins increased by 114 percent over four decades from 1970 to 2010. Additionally, while economic losses due to river floods have increased over the last 50 years, the number of casualties has decreased. The reduction in loss of life has been associated with the integration of early warning systems with emergency preparedness and planning at local and national levels (Golnaraghi et al., 2009, Kundzewicz et al., 2012).

Global early warning systems are needed to improve international disaster management. These systems can be used for both early forecasting, for better preparedness, and early detection, and for an effective response and crisis management. Their necessity was emphasized in 2005, and since then, it has been a key element of international initiatives such as the “Hyogo Framework for Action 2005-2015” and, on a continental level, the European Commission Flood Action Programme. After the 2002 floods on the Elbe and Danube rivers, the Commission supported the development of the European Flood Awareness System (EFAS) (Bartholmes et al., 2009; Thielen et al., 2009) by the Joint Research Centre to increase preparedness for riverine floods across Europe. Currently, a number of organisations are involved in rapid mapping activities after major (flood) disasters such as UNOSAT (2013), GDACS (2013), "Space and Major Disasters" (Disaster Charter, 2014), the Committee on Earth Observation Satellites (CEOS) Flood Pilotand the online Dartmouth Flood Observatory (http://floodobservatory.colorado.edu/). In Europe, Copernicus is the Earth Observation Programme which actively supports the use of satellite technology in disaster management and early warning systems for improved emergency management.

Flood warning systems typically rely on forecasts from national meteorological services and in situ observations from hydrological gauging stations. However, this capacity is not equally developed across the globe, and is highly limited in flood-prone, developing countries. Ground based hydro-meteorological observations are often either scarce or, in cases of transboundary rivers, data sharing among the riparian nations can be limited or absent. Therefore, satellite monitoring systems and global flood forecasting systems are a needed alternative source of information for national flood authorities not in the position to build up an adequate measuring network and early warning system. In recent years, there has been a notable development in the monitoring of floods using satellite remote sensing and meteorological and hydrological modelling (Schumann, et al. 2009).

A variety of satellite-based monitoring systems measure characteristics of the Earth's surface, including terrestrial surface water, over large areas on a regular basis (van Westen, 2013). Such remote sensing is based on surface electromagnetic reflectance or radiance in the optical, infrared and microwave bands. Some key advantages of microwave sensors is that they provide near-daily basis global coverage and, at selected frequencies, relatively little interference from cloud cover. Two presently-operating microwave remote sensors with near-global coverage are the Tropical Rainfall Measuring Mission[1] (TRMM) operational from 1998 to present and the Advanced Microwave Scanning Radiometer for Earth Observation System[2] (AMSR-E) which was active from June 2002 to October 2011, followed by AMSR2 which was launched in May 2012 and is onboard the Japanese satellite GCOM-W1[3], and from which, brightness temperature data are being distributed from January 2013 onwards. For future work, the European Space Agency (ESA) and NASA have other missions to put similar instruments in orbit, capturing passive microwave energy

[1] http://trmm.gsfc.nasa.gov
[2] http://aqua.nasa.gov/about/instrument_amsr.php
[3] http://suzaku.eorc.jaxa.jp/GCOM_W/w_amsr2/whats_amsr2.html

at 36.5 GHz, such as ESA’s Sentinel-3 satellites (planned launch in 2015 and 2016) and NASA’s Global Precipitation Mission (GPM) (launched in February 2014) to replace TRMM.

Using AMSR-E data initially, De Groeve et al. (2006) implemented a method for detecting major floods on a global scale, based on the surface water extent measured using passive microwave sensing. Also, Brakenridge et al. (2005, 2007) demonstrated that orbital remote sensing can be used to monitor river discharge changes. However, as underlined by Brakenridge et al. (2012, 2013), extracting the microwave signal and converting it into discharge measurements is not straight-forward and depends on factors such as sensor calibration characteristics and perturbation of the signal by land surface changes. These changes can be found for example in irrigated agricultural zones and in areas where rivers flow along forested floodplains (Brakenridge et al., 2013). As rivers discharge increases, river level (stage), river width, and river flow velocity all increase as well, and the challenge is to measure one or more of these accurately enough to provide a reliable discharge estimator, and compare against a background of other surface changes that may affect what is measured from orbit.

There remains also the need to convert such discharge estimators to actual discharge units. Using ground discharge data or climate-driven runoff models for calibration and validation, methods to convert the remote sensing signal to river discharge have been previously tested at particular stations with output from the Global Flood Detection System (GFDS, http://www.gdacs.org/flooddetection/) and by different investigators (Brakenridge et al. 2007, Brakenridge et al. 2012, Khan et al. 2012, Kugler and De Groeve, 2007, Moffitt et al. 2011, Hirpa et al., 2013, Zhang et al. 2013). Yet the results are from different approaches and not easily comparable, making an assessment of the potential performance on a global scale difficult. Furthermore, definite conclusions about the influence of various environmental factors on the signal performance have not been reached. Therefore, in this study, a rigorous broad assessment of the method is undertaken with a systematic evaluation of the relationship between skills obtained between ground- and satellite-based discharges, and the local characteristics of the stations. Specifically this study addresses mean observed discharges, river widths, land cover types, leaf area indices, climatic regions, and flood hazard maps, and the presence or absence of large floodplains, wetlands, river ice and hydraulic control infrastructure.

Our goal is to assess the potentials and limitations of the satellite-based surface water extent signal data for river discharge measurements with a large number of stations. Moreover, the relationship between ground and satellite sets of discharge measurements and the local surface characteristics is examined in order to provide guidelines for selection of observation sites. For this purpose, river catchments located in a range of different climatic and land cover types were selected in Africa, Asia, Europe, North America and South America. The remainder of the paper is structured as follows: section 2 presents the study regions and data, section 3 describes the analysis methodologies, and the results are discussed in section 4.

2 Study regions and data

2.1. Study Regions and in situ discharge data

Figure 1 shows the study basins and in situ discharge locations. The selected stations are all located near major rivers of the world (Global Runoff Data Centre, 2007). The continental distribution and the upstream catchment area of the stations are summarized in Table 1. We selected the locations to be representative of a broad variety of local conditions: they belong to nine different main land cover classes (aggregated from GlobCover, 2009) and five main types of climate (Peel et al., 2007). The characteristics are listed in Table 2.

For Africa, Asia, Europe, North America and South America, daily in situ discharge values were used from the Global Runoff Data Centre (GRDC) database. In addition, for the South African stations, the discharge data were provided by the South African Water Affairs (DWA, http://www.dwa.gov.za/). The selected stations for all these continents include daily data between 1998 and 2010, however not all stations have continuous data during this time period. From 1998, the length of the time series was required to be above six years. The longest time series available was of 13 years, with a median value of 8.5 years. In situ discharge information may itself be affected by large and variable uncertainty, mostly on the measurement of the cross-sectional area of the channel and mean flow velocity at the gauge or control site (Pelletier, 1988). Although generally unknown, these values are typically between the 5-20% at the 95% confidence levels as highlighted in studies such as Hirsch and Costa (2004), Di Baldassarre and Montanari, (2009), Le Coz et al. (2014), and Tominsk (2014). However, the uncertainty of river discharge is even higher during floods events when the stage-discharge relationship, the so-called rating curve, is used. As evaluated by Pappenberger et al. (2006), the analysis of rating curve uncertainties leads to an uncertainty of the input of 18–25% at peak discharge. Di Baldassarre and Montanari (2009) showed that the total rating curve errors increase, when the river discharge increases and varies from 1.8% to 38.4% with a mean value of 21.2%. For the purposes here, these data are, however, regarded as "ground truth". We acknowledge the possible errors, however, and note that, for some river reaches, satellite-based methods may actually track discharge changes more accurately than ground-based measurements using stage; the extent to which this is true needs to be fully investigates however.

(INSERT FIG 1 HERE)

(INSERT TABLE 1 HERE)

(INSERT TABLE 2 HERE)

2.2. Satellite-derived data

The Global Flood Detection System (GFDS) produces near real time maps and alerts for major floods using satellite-based passive microwave observations of surface water extent and floodplains. It is developed and maintained at the European Commission Joint Research Centre (JRC) in collaboration with the Dartmouth Flood Observatory (DFO). The surface water extent detection methodology using satellite-based microwave data is explained in Brakenridge et al. (2007) and Kugler and De Groeve (2007). Here, only the basic principles are recalled.

At each pixel, the method uses the difference in brightness temperature, at a frequency of 36.5 GHz, between water and land surface to detect the proportion of within-pixel water and land. The retrieved brightness temperature data are first gridded into a product with a pixel size of (near the equator) 10 x 10 km (0.09 degree x 0.09 degree), and the system provides a daily output. For our work, the merged TRMM/AMRS-E product was used (http://www.gdacs.org/flooddetection/download.aspx); the gridded data are being provided in the GCS WGS 1984 projection. For our period of study, 1998-2010, the merged data product was employed for the time period of its availability (June 2002-2010), whereas stand-alone TRMM data was used for the remaining time period (1998 to June 2002) and available latitudes. Note that from 2013 the system is providing the merged product TRMM/AMSR2, however this period is out of our scope.

In the GFDS system, the microwave signal (s) is defined as the ratio between the measurement over wet pixel (M) and the measurement over a 7 pixel x 7 pixel array of background calibration (C) pixel, known as the M/C ratio(Brakenridge et al. 2012, De Groeve, 2010). Better discharge signal values may be achieved when the measurement pixel is centred over a river reach and no hydraulic structures are present (Moffitt, et al., 2011). However, this is sometimes difficult to achieve due to the desired co-location with gauging stations (Brakenridge et al. 2012) or because the potential measurement pixels within the raster are fixed, geographically.

2.3. Other important datasets and maps

The quality of the microwave signal detected by the satellite sensors can be influenced by local ground conditions including extreme rainfall, snow/ice, land cover/use and topography (Brakenridge et al., 2012). For example, forest is a type of land cover which influences the microwave emission properties due to the biometric features of vegetation such as crown water content and shape and size of leaves (Chukhlantsev, 2006). In this study, the effects of the local ground conditions on the performance of the satellite signal were analysed as a function of the following factors:

- **River width**: channel width from Yamazaki et al. (2014), estimation based on SRTM Water Body Database and the HydroSHEDS flow direction map and for which the map was upscaled from 0.025 to 0.1 degree, taking the mean of the river grid values in the 4 x 4 area.

- **Mean observed discharge**: For each station, a mean discharge value for the study period was calculated from daily ground data (mainly from the GRDC dataset).

- **Upstream catchment area** (GRDC 2007) data: The GRDC river network was used to visually select those stations located close to the “main rivers” classified by GRDC, and to use the values of the upstream catchment area for each station. Note that upstream catchment area values are missing from all South African stations from DWA data provider.

- **Presence of Floodplains, Flooded Forest and Wetlands**: This was obtained from the Global Lakes and Wetlands Database Level 3, a global raster map at 30-second resolution which comprises lakes, reservoirs, rivers and different wetland types (Lehner and Doll, 2004).

- **Flood extent**: We used the fractional coverage of potential flooding of 25 km by 25 km cells for a 100 year return period from the Global Flood Hazard Map derived using a model grid (HTESSEL+CaMa-Flood) (Pappenberger et al. 2012).

- **Land cover**: We used land cover data from the Global Land Cover 2009 (GlobCover 2009) (ESA and UCLouvain 2010). The 19 labels were aggregated into 8 types of land cover depending on the vegetation type and density to synthesize the outputs (see Appendix Table A 1). Further visual category checking was performed using GoogleMaps display for the sites, and where necessary, land cover classes changed accordingly. An additional category was added, for sparse vegetation areas where crops are grown along or near the river channels.

- **Leaf Area Index**: A global reprocessed Leaf Area Index (LAI) from SPOT-VGT is available for a period of 1999- 2007 (http://wdc.dlr.de/data_products/SURFACE/LAI/). This LAI product is a global dataset of 36 ten-day composites at a spatial resolution of the CYCLOPES products (1 km). For our analysis, a modified version of this product was used, which was upscaled to a spatial resolution of 10 km.

- **Climatic areas:** We used the Köppen-Geiger climate map of the world (Peel et al. 2007) to distinguish the main climate areas: tropical, arid, temperate, cold and polar (see Table 2).

- **Presence of river ice:** Through the signal, the presence of river ice cover can also be detected in cold land regions. The Circum-Arctic Map of Permafrost and Ground-Ice Conditions (Brown et al., 2002) map was used here. Examples of these rivers are Yukon and Mackezie in North America and Lena River in Russia. As is the case on the ground, discharge under ice cover is left largely unmeasured as both water area and stage no longer are responsive to discharge variation.

- **Dam location**: Hydraulic structures can disrupt the natural flow of water, and therefore may alter the expected performance of the satellite signal on that location. For this analysis the Global Reservoir and Dam (GRanD) (Lehner et al., 2008) dataset was used.

3 Methodology

3.1. Satellite signal extraction

In total, 398 locations for satellite-based measurement were selected which overlap spatially and temporally with available in situ stations providing daily measurements. Since satellites never pass directly over the same track at exactly the same time, the operational GFDS applies a four day forward-running mean to systematically calculate the signal; this also commonly fills between any missing days (Kugler and De Groeve, 2007). Furthermore, for each observation site, on the GFDS system the signal is calculated as the average signal of all measurement pixels under observation for each location (which can be one or more pixels) (GDACS, 2014). Thus, in some cases, even a 10 km pixel is not large enough as a measurement site, and would entirely saturate with water during flooding. An array of measurement pixels is instead used. In this analysis, we used the signal values from the single pixels which contain the ground station, as well as a multiple pixels selection. This includes, for each location, the pixel itself and also the three nearest neighbours of the 10 x 10 km grid. In case of multiple pixels, the signal value was calculated for the spatial median, average and maxima. Similar results were obtained globally when comparing the extracted signals (single or multiple pixels) with the in situ discharge observations. Therefore, we used the temporal and spatial averaging on the multiple pixel array as in the operational GFDS. For each site, a visual check with Google maps was carried out to assure that the largest river section was included within the finalized measurement sites (see Figure 2).

(INSERT FIGURE 2 HERE)

3.2. Satellite signal calibration and validation

For those co-located ground stations and satellite measurement sites where both sets of data (signal and in situ discharge) were above six years in length, calibration and validation was performed using the ground information as reference. Several stations, mainly in North America, located close to man-made infrastructures such as weirs and generating stations were excluded from this analysis due to the rapidly changing behaviour of the in situ observed discharge. Also, in a satellite-based approach to measure river discharge, the local river characteristics and floodplain channel geometry control the accuracy of rating curves as is the case for gauging stations on the ground (Brakenridge et al., 2012, Khan et al., 2012 and Moffitt et al. 2011). Thus we expect some measurement sites to exhibit a more robust response to discharge changes, and a higher signal to noise ratio, than others.

It has been acknowledged that for large rivers, using the daily GFDS signal as a floodplain flow surface area indicator of discharge might result in a few days lag when comparing with ground-based discharge (Brakenridge, 2013). Thus, stage may immediately rise at a gauging station as a flood wave approaches, but flow expansion out into the floodplain requires some increment of time. This time lag may introduce error into the scatterplots used to calculate the rating equations, and therefore lower skill scores obtained when analysing both datasets. In addition, in previous studies (Khan et al. 2012, Zhang et al. 2013), it was observed that, in some cases, an overestimation of satellite measured discharge existed during low flow periods when using a single rating equation for the full period to calibrate signal into discharge units. For this reason, we decided to use a rating equation for each month individually, and grouping daily into monthly data. In this case the time series data for a fixed month can be treated as stationary and the derived daily discharge values adjusted better also during low flow periods.

To calibrate satellite signal into discharge measurements, the first five years of data were used for both satellite signal and ground discharge for each location. Regression equations were obtained using monthly means from daily values and with which GFDS measured discharge was derived.

$$Q_{\text{GFDSmeasured of X month}} = a_{\text{month}} + b_{\text{month}} * \text{signal} \qquad (1)$$

For the sake of simplicity, for this paper, the equations were restrict to linear equations. However, as the relation is purely empirical, we leave for follow on-work more research on flexible way to fit these relations. Note that fitting straight lines to curves will reduce goodness of fit and predictive accuracy. Power law fitting was also tested to calibrate the signal into discharge units yielding similar results (see Open Discussion Author's Response).

The validation of the satellite derived daily discharge data was carried out with daily in situ data on a two-year period, and skills scores were calculated to quantify the agreement between both satellite and ground measured discharge. We are aware of the limited number of years (data) with available time series for both variables, which might influence the robustness of the calibration. In some cases there were longer time series available, but to standardised the analysis for all the stations we used five years (1998-2002 or 2003-2008 for Northern stations with AMSR-E signal) and the following two years for validation purposes (2003-2004 and 2009-2010 respectively). Note that for 36 out of the 322 stations available data length was between six years and three months to almost seven years. Validation was still carried out for the same period, but the data used for calibration was slightly reduced. As an example, Figure 3a presents the scatterplot for the month of March for the Senanga Station (Long 23.25, Lat. -16.116) in the Zambezi River (Africa) with mean values derived from the period 1998 to 2002. For the same location, Figure 3b shows the in situ observed and the GFDS measured discharge derived from the GFDS signal for the period 2003-2004.

(INSERT FIGURE 3 HERE)

3.3. Skill scores

The initial analysis of the correlation of the remote sensing signal to in situ discharge was assessed for each station and site pair through the Pearson correlation coefficient (R). For the validation, the performance of the satellite-measured discharge was also assessed using the Nash-Sutcliffe Efficiency (NSE) statistic in addition to the R skill score. Spearman's rank correlation coefficient (ρ) was also calculated to assess the validation performance.

One of the advantages of the R coefficient is its independence on the units of measurement, which permits the comparison of dimensionless GFDS signal data. A small value indicates a weak or non-linear relationship between the satellite signal and discharge. For this study, we grouped the computed R values into three ranges as follows: <0.3, [0.3-0.7], and >0.7. While Pearson benchmarks linear relationship, Spearman benchmarks monotonic relationship. Spearman's validation scores just obtained a mean value 6% higher than Pearson mean score (see Open Discussion Author's Response). On this manuscript, results are analysed based on the scores obtained using Pearson correlation coefficient.

Nash-Sutcliffe efficiency (NSE) (Nash and Sutcliffe, 1970) is typically used to assess the predictive power of hydrological models and was here calculated to describe the accuracy of satellite-derived discharge in comparison to gauge-observed discharge values. Higher values of the Nash-Sutcliffe statistic should indicate more correlated results, without other factors taken into account, such as autocorrelation (Brakenridge et al., 2012). However, the degree of correlation of these variables does not verify the discharge magnitudes (Brakenridge et al., 2013). A NSE value of 1 corresponds to a perfect match of modelled to the observed data whereas NSE = 0 indicates that the model predictions are as accurate as the mean of the observed data. The resulting scores will be classified as in Zaraj, et al. (2013): < 0, [0.2-0.5], [0.5-0.75], and > 0.75.

3.4 Factors affecting the satellite signal

Understanding the influence of local factors on the accuracy of the satellite flood detection is critical for practical use of the remotely sensed signal. We analysed the accuracy effects of river width, mean daily discharge, upstream catchment area, presence of large floodplain, flooded forest and wetlands, the potential flood extent, land cover type, Leaf Area Index (LAI), climatic areas, presence of river ice and hydraulic structures. To assess their influence, the fractional coverage over the measurement site was retrieved for variables with spatial coverage.

First, we use the skill scores (R and NSE) obtained from a simple analysis for each individual factor or variable. Second, we seek to understand which of the surface variables have the greatest importance in determining sites with a good or poor performance. For this purpose, we use a decision tree technique called Random Forest (RF). Among other features, this allows ranking of the relative importance of each variable. The technique is described by Breiman (2001) and implemented in R by Liaw and Wiener (2002), where the reader is referred for a more detailed explanation. As a summary of the Random Forest algorithm, *ntree* bootstrap samples are randomly selected from the data set, a different subset is used for each bootstrap and for each sample a tree is grown, obtaining *ntree* trees. Random Forest is called an ensemble method because it applies the method for a number of decision trees, in this case 500, in order to improve the classification rate. Some stations are left out of the sample (out-of-bag) and used to gain an internal unbiased estimate of the generalisation error (oob errors) and to obtain estimates of the importance of the variables (Breiman, 2001). These values are averaged over the *ntree* trees. For the variables classification, the node impurity is measured by the Gini index. Gini´s mean difference was first introduced by Corrado Gini in 1912 as an alternative measure of variability and the parameters derived from it, such as the Gini index, also referred to as the concentration ratio (Yitzhaki and Schechtman,2013). The Gini index is mostly popular in economics, however it is also used in other areas, such as building decision trees in statistics to measure the purity of possible child nodes, and it has been compared with other equality measures (Gonzalez,L., et al. 2010).. The variables with higher decrease in Gini values (lower Gini) are those with higher importance on the classification analysis.

Although for “black-box models” such as Random Forest the information is hidden inside the model structure, the prediction power is high (Palczewska et al., 2013). This method is relatively robust given outliers and noise because it uses randomly chosen subsets of variables at each split of each tree (Breiman, 2001; Chan et al., 2008). To further increase robustness, Strobl et al. (2009) states that results from the random forest and conditional variable importance should always be tested by doing multiple random forest runs using different seeds and sufficiently large ntree values to obtain robust and stable results.

The quality index chosen to rank variable importance and classify good or poor locations, in the Random Forest analysis, was the Nash-Sutcliffe Efficiency (NSE) score. A threshold of NSE=0 splits the data into two groups, obtaining about 50% of the data above (true or good predictive) and below (false or poor predictive) that value of NSE. The results presented here are the average of 200 runs. Furthermore, four different training sets were used by a random 70%/75%/80%/90% of the stations and were validated with the remaining 30%/25%/20%/10% of stations, respectively.

4 Results and discussion

As a first step we analysed the relationship between the satellite signal and the in situ observed discharge to have an initial understanding of the performance between the two datasets (Section 4.1). Then we calibrated the satellite signal with in situ discharge data. With the regression equations obtained, we calculated satellites discharge measurements. A two-year validation period was carried out for each station using the skill scores as described in Section 3.3 (Section 4.2). This was followed by an assessment for how different variables contribute in a positive or negative way to the overall skill (Section 4.3). Variables included in the analysis are daily mean river discharge, river width, upstream catchment area, potential flood hazard area, land cover, leaf area index, climatic zones, presence of large floodplains, flooded forest and wetlands, river ice and hydrologic structure. Finally, the relative importance of all variables in comparison to each other has been assessed (Section 4.4).

Before analysing the validation results, it is important to highlight two possible different sources of error which might influence the outputs. Firstly, the signal to noise ratio might be low for a site or have intermittent instrument noise occasionally producing positive spikes in discharge. Secondly, the rating curve may be offset, which will result in a consistent bias on the discharge values for that location even though the time series are strongly correlated.

4.1. Correlation of raw satellite data vs. gauge observations

The first step was to look at the "raw" correlation between daily ground station-measured water discharge and the satellite signal and to calculate the empirical linear relation between these two variables for each site. The full time series, including low flows, were used for the calculation and executed for 398 stations. Figure 4 shows the R skills obtained. 169 out of 398 sites have an R > 0.3 and 42 of them have R>0.5. Perhaps, correlations might have been higher if regression would have not been restricted to linear equations (Brakenridge et al., 2007, 2012).

(INSERT FIGURE 4 HERE)

4.2. Satellite signal calibration, validation and evaluation through skill scores

For the stations with over six years of contemporary data for both in situ discharge and satellite signal, we obtained regression equations for each month of the year and station using the first five years of data. Next, using these equations we carry out a calibration of the daily signal into discharge units. Afterwards, the validation of the GFDS measured discharge was implemented for the following two years. In some regions such as Northern Asia, the lack of available recent long time series (after 2002) meant that the number of stations available for calibrating the satellite into discharge measurements was reduced. Stations where the number of years matching observed discharge and satellite signal was shorter than six years were excluded from the validation exercise despite performing well. Finally, out of 398 a total of 332 stations remained for calibration and validation.

Figure 5 shows that for NSE score, 154 out of 332 stations are larger than 0; 13 located in Africa, 77 in North America, 62 in South America, 1 in Asia and 1 in Europe. Nevertheless, it needs to be noted that in arid regions, results calculated with the skill scores such as NSE are penalised, by low average discharge compared to high flow conditions. If instead of using all the available time series, a "dry stream" threshold would have been applied, the scores obtained for these sites could have been higher when analysing the remaining dataset period where flow is present.

(INSERT FIGURE 5 HERE)

4.3. Analysis of the factors affecting the satellite signal

4.3.1. River width and presence of floodplain and wetlands.

As a first step to analyse the potential relationship between the individual local characteristics and the performance of the locations in global terms, we study the R score of the validation for the 322 stations in relation with the maximum river width value at each location (Figure 6a). Results indicate that locations with a river width higher than 1 km are more likely to score an R larger than 0.3. In fact, the mean R score is 0.60. Where 26 out of 64 (~41%) have R> 0.75. However, there is a number of stations with lower river width that also obtained high scores. As the retrieval of the satellite signal also depends on the floodplain geometry. As soon as the river floods and water goes over-bank, the proportion of water in the wet pixel greatly increases. So the score should be also high for small rivers with a proportionally big floodplain. Figure 6b shows the R scores by locations where the majority of the area belongs to floodplain, flooded forest and wetlands category or, their absence. In our study, higher median scores were obtained for those located in large freshwater marsh and floodplains, followed by those on swamps and flooded forest. These results give a first indication on the characteristics of the locations with better performance.

(INSERT FIGURE 6 HERE)

4.3.2. River discharge and potential flooding

Flooding is determined by the discharge as well as the potential flood hazard. Figure 7a shows that 84 out of 95 stations with R<0.3, also have mean discharge values lower than 500 m^3s^{-1} (Log10

(500) ≈2.7), of which 55 stations in fact had a mean discharge lower than 200 m^3s^{-1}. These stations are mainly located in South Africa, and in some areas of North America. Therefore, it can be concluded that the mean discharge can be considered a key variable that determines the appropriateness of locations for which satellite discharges can be derived: As 77% of the stations with Q<500 m^3/s, have R< 0.3, while 91.5% of the stations with Q>500 m^3/s have R >0.3, locations with discharge of less than 500 m^3s^{-1} might not provide reliable results for a global satellite-based monitoring system. Alternatively, non-permanent rivers and streams exhibiting only seasonal or ephemeral flow (typical for dry regions) may require a different monitoring approach, wherein a "dry" threshold is established for the signal data.

After excluding the global stations with low skill score due to low flows and studying the remaining stations, we can better understand the performance of the system in relation to other local characteristics. Figure 7b shows for each location the relationship between the validation R and the percentage of area in each pixel covered by potential flooding during a 100 year return period flood event, obtained with the model grid (HTESSEL+CaMa-Flood) (downscaled from a 25 x 25 km pixel, Pappenberger et al., 2012). 100 means totally flooded across its area, 50 means 50 % of the area within the cells is flooded, and 0 means that the area is not flooded. Although there is not a clear trend for all the points, result indicate that locations with a percentage of potential flooding larger than 40%, are expected to score an R larger than 0.3.

(INSERT FIGURE 7 HERE)

4.3.3. Land cover types and climatic areas

Figure 8 presents a global evaluation of the R score obtained during the validation and its classification by the land cover type of the stations. The bare land cover category was excluded from this study as only one of the selected locations belong to that class. Looking at the median of the boxplot (see Figure 8), we found that some of the locations with higher density of vegetation such as those located on "closed forest" and "mosaic with predominant vegetation" (included forest, scrublands and grasslands) obtained lower median scores values. In contrast, the locations with lower vegetation density such as "sparse vegetation", "mosaics with predominant cropland/grasslands", "open forest" and "closed to open forest" land cover types obtained larger median R scores, around 0.6-0.8. Similar results can be observed when looking at the interquartile range or spread of the boxplots: "closed to open forest" and "mosaics with predominant cropland/grasslands" obtained better results. Meanwhile, "closed forest" and "mosaic with predominant vegetation" had lowers scores. In addition, those sites with a combination of sparse vegetation and crops growing near the river channel had a lower median value where comparing with those on sparse or mosaic crops land cover. Note that the sites with "sparse with crops" are located in arid climatic areas, whereas most of the "sparse" are in cold or polar regions, therefore run by different processes. In addition, sites with a majority of artificial/urban land cover (not shown) obtained a low median value of 0.267.

(INSERT FIGURE 8 HERE)

The relationship between locations by main Köppen-Geiger climatic areas (Peel et al. 2007) and R score obtained is shown in Figure 9. Globally the tropical regions (Africa and South America) obtained the highest median scores (R≈0.8), followed by cold regions (R≈0.6). Lower median score values (R≈0.3) were obtained for arid and temperate regions. It is important to clarify that these results are not only due to direct climate characteristics but also for example due to the characteristics of the rivers on those areas. In the case of the arid regions, it is mainly related with reduce daily average discharges, a characteristic of many of these stations. Note that polar climate was excluded from this evaluation as only three locations belong to that class.

(INSERT FIGURE 9 HERE)

4.3.4. Leaf Area Index (LAI)

Leaf Area Index (LAI) values typically range from 0 for bare ground to 6 or above for a dense forest, however CYCLOPES underestimates over dense vegetation (forest) (Zhu et al., 2013). Therefore, for this product LAI range is limited to [0-4], as seen in our analysis. Despite this, CYCLOPES is the most similar product to LAI references map (*Ibid.*). According to the study carry out by Zhu et al. (2013) monthly CYCLOPES LAI values for the period 1999 to 2007 by four main groups of vegetation are predominantly as follows: bare ground [0], forest [0-3.5], other woody vegetation [0-1.5], herbaceous vegetation [0-2], and cropland/natural vegetation mosaics [0-3]. The highest annual mean LAI values are obtained by evergreen broadleaf forest (3.16), included in our "closed to open forest" class.

We decided to study the relationship between the mean Leaf Area Index and the skill obtained in the validation for each location, also looking at complementary variables such us the land cover and the geographical region which the stations belong to. Figure 10 shows that locations with a mean [LAI > 2] predominantly have a "closed to open forest" type in South America (31 stations) of which 29 have an R score higher than 0.6. For [LAI > 2] there is also 12 North American locations with "closed forest" land cover but in general with poorer scores for those locations. Additionally, 18 stations with mosaic vegetation from North and South America obtained [LAI > 2] and 16 out of them, a [R>0.6]. For [LAI < 2], both the land cover and geographical locations are distributed along the scatterplots, from poor to high correlations.

(INSERT FIGURE 10 HERE)

4.3.5. River ice

Figure 11a shows the scores obtained for the locations with presence or absence of river ice, including a range from continuous to sporadic (Brown et al., 2002). It can be seen that stations located in areas with river ice tend to have a good correlation between in situ and satellite measured discharge (based on 33 stations), as the system tends to capture well the annual spring ice break-up and freezing as indicated in the study by Brakenridge et al.(2007) and Kugler (2012). At these locations, once ice-covered there is no sensing capability from the system: which may seems analogous to low flow conditions, and for which sites we obtained lower scores. However, there is an important difference when analysing time series of signal between ice covered high latitude river and all-year-around low flow rivers. When on the sites with river ice melting process takes place, there is an increase of runoff happening and for many places the signal strongly indicates this increased flow. On the other type of rivers, low flows is generally a characteristic for most of the year and if the signal to noise is low, the signal retrieved is very noisy: one motivation for setting a "dry" threshold for such sites.

4.3.6. Hydraulic structures

The correlation between satellite and discharge data depends on both variables. Typically it is assumed that observed discharges are "ground truth", however, when influenced by structures and dams the ground discharge may not be well-monitored by flow area/flow width variation. For example, when there is a major increase in river discharge but a flood is avoided by artificial levees, we cannot expect that the satellite signal will accurately capture the flood hydrograph; as well, downstream flooding may be attenuated by an upstream flood control dam and reservoir; so that the gauge location is critical. Figure 11b shows the influence of the presence or absence of a nearby dam using the Global Reservoir and Dam (GRanD) database (Lehner et al., 2008) or visually identified hydraulic control infrastructure. Locations where the dam or other element was present (48 stations) obtained lower median R score. Therefore, ideally, observation sites should be located in areas without hydraulic control infrastructures.

(INSERT FIGURE 11 HERE)

4.4. Variable importance

Based on the individual analysis of the signal potential influence factors we found that to understand the site performances, on some occasions multiple variables need to be analysed in a simultaneous way. For example, the general low scores obtained on the Eastern USA stations might be due to a number of factors: ~64% of these stations have a mean discharge value lower than 500 m^3s^{-1} and ~88% of the stations are located at river width lower than 1km. In addition, ~59% of the stations are located in wetlands areas. Another example, in this case regarding the exceptions of the low R and mean observed discharge higher than 500 m^3s^{-1}, all the 11 locations have a potential probability of flooding lower than 21%, the land cover of 10 out of 11 is forest, 5 of them located in wetlands and two of them have a nearby hydraulic structure. Despite exhibiting a mean discharge greater than 500 m^3s^{-1}, these other local characteristics may be the cause of the poor performance. Therefore, we decided to use a classification decision tree technique (Random Forest), which split the dataset at each node according to the value of one variable at a time (the best split) from a selected set of variables to understand the importance of each variable. Random Forest is called an ensemble method because it is performed for a number of decision trees, in this case 500 trees, in order to improve the classification rate.

The result presented here is the rank of the importance of variables to classify a location with a good or poor performance. These values are obtained as an output of the Random Forest analysis and are, in addition, the average of 200 independent runs. As explained in section 3.4 the variable importance based on the mean decrease in Gini index was calculated for the Nash-Sutcliffe Efficiency (NSE) score obtained from the validation. We used a NSE=0 to distinguish the sites with a good (above 0) from poor performance (below 0) and we also tested it with a threshold NSE of 0.50.

Figure 12 presents the variable importance for the four test groups. Features which produced large values of the “Mean Decrease in Gini” are ranked as more important than features which produced small values. For our locations and data available the mean daily observed discharge has the highest importance, followed by the climatic region, land cover / mean LAI and upstream catchment area. Meanwhile, the presence of hydraulic structures (mainly dams) and of river ice has the lowest importance to classify a location as good or poor performance. However, this does not mean that it has no influence. Although discharge is correlated with upstream catchment area and at some degree also leaf area index with land cover type, both were included in this case to understand which variable might help us most to classify the sites.

Although, the effect of the correlations on these measures has been studied recently (see Archer and Kimes (2008), Strobl et al. (2008), Nicodemus and Malley (2009), Nicodemus et al. (2010), Nicodemus (2011), Auret and Aldrich (2011), Tolosi and Lengauer (2011), Grömping, U. (2009) and Gregorutti et al. (2013)) there is no yet a consensus on the interpretation of the importance measures when the predictors are correlated and on what is the effect of this correlation on the importance measure.

In order to test the effect on the results when correlated variables were included in our analysis, an independent Random Forest analysis was carried out (not shown in the paper) for the same variables but excluding the river width and the presence of floodplains and wetlands variables. Results also showed that the mean daily observed discharge had the highest importance and the presence of hydraulic structures (mainly dams) and of river ice had the lowest importance to classify a location as good or poor performance.

(INSERT FIGURE 12 HERE)

5 Conclusions and future research

In this article we presented an evaluation of the skill of the Global Flood Detection System to measure river discharge from remote sensing signal. From the 322 stations validated the average continental R skills are as follow: Africa 0.382, Asia 0.358, Europe 0.508, North America 0.451 and South America 0.694. Approximately 48% of these stations have an NSE score higher than zero; 13 located in Africa, 77 in North America, 62 in South America, 1 in Asia and 1 in Europe. Results showed that the majority of the stations that received low skills scores, were due to low flow conditions. For example, 84 out of 95 stations with $R<0.3$, have mean discharge values lower than 500 m^3s^{-1}. These are located mainly in South Africa with 25 cases and North America with 53 cases, which penalised their average continental skills. Note that our focus was on factors affecting the method, globally, and that these skill values do not directly indicate at-a-site measurement accuracy (which could be improved, for example, by use of non-linear rating equations and/or accommodation of any phase shift or timing differences in flow area- versus state-based discharge monitoring).

In order to better understand the impact of the local conditions on the performance of the sites, we looked first at specific factors individually. In general terms, higher skill scores were obtained for location with one or more than one of the following characteristics: a river width higher than 1km, a large floodplain area, in flooded forest, with a potential flooded area per pixel greater than 40% during a 100 year return period flood event, a land cover type of sparse vegetation, croplands or grasslands and closed to open and open forest, Leaf Area Index above 2, located in a tropical climatic area, and where no dams or hydraulic infrastructures are present. Also, out of our locations, high latitude rivers with seasonal ice-cover tend to exhibit good performance.

Secondly, we performed a classification decision tree analysis, based on Random Forest, to obtain the variable importance when classifying a site as good or poor. The output of this analysis showed that mean observed discharge, climatic region, land cover and mean leaf area index (LAI) and upstream catchment area and were the variables with higher importance, whereas river ice and dam obtained the lowest importance. Both the individual and the combined classification analysis of these local characteristics give us critical evidence of the relationship between the ground and satellite discharge measurements and when it is expected to perform well. Furthermore, it provides a guideline for future selection of measuring sites.

The locations with a very good performance will be selected for a potential future project where satellite measure discharge could be calculated for longer periods and on a daily basis from the remote sensing signal, analogous to the Dartmouth Flood Observatory method. This will represent a major step forward in developing continental and global hydrological monitoring systems as these data can fill the gaps where real time ground discharge measurements are not available (the case at many locations globally). We found that some of the sites with good performance are located within international river basins such as the Niger, Volta and Zambezi in Africa. In addition, for the studied locations with good signal performance but rather short contemporary time series with in situ observed discharge (such as the Siberian stations), the calibration of the signal to obtain discharge measurements could be executed at any point when additional ground data is available. This will also be beneficial for all stations including those with time series above seven years long.

Zhang et al. (2013) recently demonstrated the potential of integrating satellite signal provided by the Global Flood Detection System in improving flood forecasting. This first attempt of data assimilation was carried out for a single station (Rundu, northern Namibia- included in this study) with the conceptually simple Hydrological MODel (HyMOD). Hence, a prospective study with the inclusion of all these stations for post-processing through data assimilation and error correction of the stream-flow forecast in hydrological models could be done. For instance, for the pre-operational Global Flood Awareness System (GloFAS) (Alfieri et al. 2012) and the African Flood Forecasting System (AFFS) (Thiemig et al. 2014) in an analogous way as it is already being done with ground gauge observed streamflow on the European Flood Awareness System (Bartholmes et al., 2009; Thielen et al., 2009). Hence, work towards the integration of global flood detection and forecasting systems such as GFDS and GloFAS, respectively, can provide a more comprehensive information for decision makers.

Acknowledgements. We acknowledge the Global Runoff Data Centre and South African Water Affairs for providing historic discharge measurements. Furthermore we would like to acknowledge the team from the Joint Research Centre Crisis Monitoring and Response Technologies (CRITECH) for support and access to the Global Flood Detection System signal historical data. Also, Philippe Roudier, Simone Russo, Angel Udias and Feyera Hirpa are thanked for their valuable input and methodology advice and Ad de Roo for PhD supervision and the editor and the two reviewers are gratefully acknowledged for their valuable feedback. G. R. Brakenridge acknowledges funding support from the NASA Hydrology Program.

References

Alfieri, L., Burek, P., Dutra, E., Krzeminski, B., Muraro, D., Thielen, J. and Pappenberger, F.: GloFAS-global ensemble streamflow forecasting and flood early warning, Hydrology and Earth System Sciences, vol. 17, no. 3, pp. 1161-1175, 2013.

Archer, K. J. and Kimes, R. V. Empirical characterization of random forest variable importance measures. Computational Statistics and Data Analysis, 52:2249–2260, 2008. doi: 10.1016/j.csda.2007.08.015

Auret, L. and Aldrich, C. Empirical comparison of tree ensemble variable importance measures. Chemometrics and Intelligent Laboratory Systems, 105:157–170, 2011. doi: 10.1016/j.chemolab.2010.12.004

Bartholmes, J.C., Thielen, J., Ramos, M.H. and Gentilini, S.: The European flood alert system EFAS - Part 2: Statistical skill assessment of probabilistic and deterministic operational forecasts. Hydrology and Earth System Sciences, 13(2): 141-153, 2009.

Brakenridge, G. R., S. V. Nghiem, E. Anderson, and S. Chien, Space-based measurement of river runoff, Eos Trans. AGU, 86(19), 185–188, 2005, doi:10.1029/2005EO190001.

Brakenridge, G.R., Nghiem, S.V., Anderson, E. & Mic, R.: Orbital microwave measurement of river discharge and ice status, Water Resources Research, vol. 43, no. 4, 2007, W04405, doi:10.1029/2006WR005238.

Brakenridge, G.R., Cohen, S., Kettner, A.J., De Groeve, T., Nghiem, S.V., Syvitski, J.P.M. and Fekete, B.M.: Calibration of satellite measurements of river discharge using a global hydrology model, Journal of Hydrology, vol. 475, pp. 123-136, 2012.

Brakenridge, G.R., De Groeve, T., Cohen, S., and Nghiem, S. V.: River Watch, Version 2: Satellite River Discharge and Runoff Measurements: Technical Summary, University of Colorado, Boulder, CO, USA, http://floodobservatory.colorado.edu/SatelliteGaugingSites/technical.html, last access: 1 December 2013.

Breiman, L.: Random Forests. Machine Learning, 45, 5–32, 2001.

Brown, J., O.J. Ferrians, Jr., J.A. Heginbottom, and E.S. Melnikov.: Circum-Arctic Map of Permafrost and Ground-Ice Conditions. Version 2. [Permafrost], Boulder, Colorado USA: National Snow and Ice Data Center, 2002.

Committee on Earth Observation Satellites (CEOS) Flood Pilot, http://www.ceos.org/, last access: 1 September 2014.

Chan, J.C.-. & Paelinckx, D.: Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery, Remote Sensing of Environment, vol. 112, no. 6, pp. 2999-3011, 2008.

Chukhlantsev, Alexander A.: Modeling of microwave emission from vegetation canopies, Microwave Radiometry of Vegetation Canopies. Springer Netherlands, Chapter 6. pp. 147–175, 2006.

Di Baldassarre, G. and Montanari, A.: Uncertainty in river discharge observations: a quantitative analysis, Hydrol. Earth Syst. Sci., 13, 913-921, doi: 10.5194/hess-13-913-2009, 2009.

De Groeve, T., Brakenridge, G. R., and Kugler., Z.: Near Real Time Flood Alerting for the Global Disaster Alert and Coordination System, eds B. Van de Walle, P. Burghardt, and C. Nieuwenhuis Proceedings of the 4th International ISCRAM Conference, 33-40, 2006.

De Groeve, T., and Riva, P.: Global Real-time Detection of Major Floods Using Passive Microwave Remote Sensing, Proceedings of the 33rd International Symposium on Remote Sensing of Environment Stresa, Italy, May 2009.

De Groeve, T.: Flood monitoring and mapping using passive microwave remote sensing in Namibia, Geomatics, Natural Hazards and Risk, 1:1, 19-35, 2010.

Di Baldassarre, G. and Montanari, A.: Uncertainty in river discharge observations: A quantitative analysis, Hydrology and Earth System Sciences, vol. 13, no. 6, pp. 913-921, 2009.

Disaster Charter, 2013. Space and Major Disasters. http://www.disasterscharter.org/, last accessed 1 September 2014.

EM-DAT, The OFDA/CRED International Disaster Database, Université Catholique de Louvain, Brussels, Belgium, http://www.emdat.be, last access: 1 December 2013.

Fekete, B.M., Vorosmarty, C.J., Grabs, W., 1999. Global, composite runoff fields based on observed river discharge and simulated water balances, GRDC Report 22, Global Runoff Data Center, Koblenz, Germany.

GDACS, Global Disaster Alert and Coordination System, Global Floods Detection System. http://www.gdacs.org/, last accessed 1 December 2013.

Global Runoff Data Centre: Major River Basins of the World, 2007. 56068 Koblenz, Germany: Federal Institute of Hydrology (BfG). http://grdc.bafg.de/, last accessed 20 January, 2013

Global Runoff Data Centre, The. River Discharge Time Series. 56068 Koblenz, Germany: Federal Institute of Hydrology (BfG). http://grdc.bafg.de/, last accessed 20 January, 2013

Golnaraghi M., J. Douris, J.-B. Migraine: Saving Lives Through Early Warning Systems and Emergency Preparedness, Risk Wise, Tudor Rose, pp 137–141, 2009.

Gonzalez, L., Velasco Morente, F., Gavilan Ruiz, J.M., Sanchez-Reyes Fernandez, J.M. The Similarity between the Square of the Coefficient of Variation and the Gini Index of a General Random Variable. Journal of Quantitative Methods for Economics and Business Administration 10: 5–18.2010, ISSN 1886-516X.

Gregorutti,B., Michel, B., Saint-Pierre, P. Correlation and variable importance in random forests. Cornell University Library, 2013. arXiv: 1310.5726 [stat]

Grömping, U. Variable Importance Assessment in Regression: Linear Regression versus Random Forest. The American Statistician. 11/2009; 63:308-319, 2009. doi: 10.1198/tast.2009.08199

Hirpa FA, Hopson TM, De Groeve T, Brakenridge GR, Gebremichael M, Restrepo PJ. Upstream satellite remote sensing for river discharge forecasting: Application to major rivers in South Asia. Remote Sens Environ, 131:140-51, 2013.

Hirsch R. M, Costa J. E.:U.S. Stream Flow Measurement and Data Dissemination Improve EOS, Transactions, American Geophysical Union. Vol. 85, No. 20, 18 May 2004, 197-203 pp, 2004.

Khan, S.I., Hong, Y., Vergara, H.J., Gourley, J.J., Robert Brakenridge, G., De Groeve, T., Flamig, Z.L., Policelli, F. & Yong, B.: Microwave satellite data for hydrologic modeling in ungauged basins, IEEE Geoscience and Remote Sensing Letters, vol. 9, no. 4, pp. 663-667, 2012.

Kugler, Z., and De Groeve, T.: The Global Flood Detection System, Office for Official Publications of the European Communities, Luxembourg, 2007.

Kugler, Z.: Remote sensing for natural hazard mitigation and climate change impact assessment, Quaterly Journal of the Hungarian Meteorological Service. Vol.116, No.1, January-March 2012, pp.21-38, 2012.

Kundzewicz, Z. W: Changes in Flood Risk in Europe, Wallingford: IAHS Press. 516 p. IAHS special publication; 10, 2012. United Nations: Report of the United Nations Conference on Sustainable, Development Rio de Janeiro, Brazil. 20–22 June 2012, A/CONF.216/16, 2012.

Le Coz, J., Renard, B., Bonnifait, L., Branger, F. & Le Boursicaud, R.: Combining hydraulic knowledge and uncertain gaugings in the estimation of hydrometric rating curves: A Bayesian approach, Journal of Hydrology, vol. 509, pp. 573-587, 2014.

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. doi: 10.1016/j.jhydrol.2004.03.028

Lehner, B., Reidy Liermann, C., Revenga, C., Vörösmarty, C., Fekete, B., Crouzet, P., Döll, P., Endejan, M., Frenken, K., Magome, J., Nilsson, C., Robertson, J., Rödel, R., Sindorf, N., Wisser, D.: High resolution mapping of the world's reservoirs and dams for sustainable river flow management, Frontiers in Ecology and the Environment. Source: GWSP Digital Water Atlas. Map 81: GRanD Database (V1.0), 2008. Last access: 11/03/2014. http://atlas.gwsp.org/index.php?option=com_content&task=view&id=209&Itemid=1

Liaw, A. and Wiener, M. Classification and Regression by randomForest. R News 2(3), 18—22, 2002.

Moffitt, C.B., F. Hossain, R.F. Adler, K.K. Yilmaz, and H.F. Pierce. Validation of a TRMM-Based Global Flood Detection System in Bangladesh. International Journal of Applied Earth Observation and Geoinformation Volume 13, Issue 2, April 2011, Pages 165-177, DOI: 10.1016/j.jag.2010.11.003.

MunichRe, Munich Reinsurance: January 2014 press release, 2014.Münchener Rückversicherungs-Gesellschaft, Geo Risks Research, NatCatSERVICE http://www.munichre.com/en/media_relations/press_releases/2014/2014_01_07_press_release.aspx, last access 20 January 2014

Nash, J.E. & Sutcliffe, J.V.: River flow forecasting through conceptual models part I - A discussion of principles Journal of Hydrology, vol. 10, no. 3, pp. 282-290, 1970.

Nicodemus, K. K. Letter to the editor: On the stability and ranking of predictors from random forest variable importance measures. Briefings in Bioinformatics, 12:369–373, 2011. doi: 10.1093/bib/bbr016

Nicodemus, K. K., Malley, J. D., Strobl, C., and Ziegler, A. The behavior of random forest permutation-based variable importance measures under predictor correlation. BMC Bioinformatics, 11:110, 2010. doi: 10.1186/1471-2105-11-110

Palczewska, A., Palczewski, J., Robinson, R.M. and Neagu, D.: Interpreting random forest models using a feature contribution method, Proceedings of the 2013 IEEE 14th International Conference on Information Reuse and Integration, IEEE IRI 2013, pp. 112, 2013.

Pappenberger, F., Matgen, P., Beven, K.J., Henry, J.B., Pfister, L., de Fraipont, P. Influence of uncertain boundary conditions and model structure on flood inundation predictions. Adv. Water Resour. 29, 1430–1449, 2006. doi: 10.1016/j.advwatres.2005.11.012

Pappenberger, F., Dutra, E., Wetterhall, F. & Cloke, H.L.: Deriving global flood hazard maps of fluvial floods through a physical model cascade, Hydrology and Earth System Sciences, vol. 16, no. 11, pp. 4143-4156, 2012.

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, doi: 10.5194/hess-11-1633-2007, 2007.

Pelletier, P.M.: Uncertainties in the single determination of river discharge: a literature review. Canadian Journal of Civil Engineering, 15:834–850, 1988.

Rosso, R. A linear approach to the influence of discharge measurement error on flood estimates. Hydrol. Sci. J. 30 (1), 137–149, 1998. doi: 10.1080/02626668509490975

Sandri, M. and Zuccolotto, P.. A bias correlation algorithm for the Gini variable importance measure in classification trees. Journal of Computational and Graphical Statistics, 17:611-628, 2008. [184], doi: 10.1198/106186008X344522

Schumann, Guy, Paul D. Bates, Matthew S. Horritt, Patrick Matgen, and Florian Pappenberger.: Progress in Integration of Remote Sensing–derived Flood Extent and Stage Data and Hydraulic Models. Reviews of Geophysics 47, RG4001, no. 4, 2009. doi: 10.1029/2008RG000274.

South African Water Affairs (DWA) database, http://www.dwa.gov.za/Hydrology/, last access: 10 July 2013.

Strobl, C., Malley, J. & Tutz, G.: An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests, Psychological methods, vol. 14, no. 4, pp. 323-348, 2009. doi: 10.1186/1471-2105-9-307

Thielen, J., Bartholmes, J., Ramos, M.-H., and de Roo, A.: The European Flood Alert System. Part 1: Concept and development, Hydrol. Earth Syst. Sci., 13, 125–140, doi: 10.5194/hess-13-125-2009, 9278, 2009.

Thiemig, V., Bisselink, B., Pappenberger, F., and Thielen, J.: A pan-African Flood Forecasting System, Hydrol. Earth Syst. Sci. Discuss., 11, 5559-5597, doi: 10.5194/hessd-11-5559-2014, 2014.

Tolosi, L. and Lengauer, T. Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics, 27:1986–1994, 2011. doi: 10.1093/bioinformatics/btr300

Tomkins, K.M.: Uncertainty in streamflow rating curves: Methods, controls and consequences. Hydrological Processes, 28(3), pp. 464-481, 2014.

UNISDR: Global Assessment Report: Revealing Risk, Redefining Development, Chapter 2.2. Global disaster risk trends, United Nations, printed in the UK, ISBN 978-92-1-132030-5, page 22-27, 2011.

UNOSAT, UNITAR Operational Satellite Applications Programme http://www.unitar.org/unosat/maps, last accessed 1 December 2013

Van Westen, C.J.: Remote sensing and GIS for natural hazards assessment and disaster risk management. In: Shroder, J. (Editor in Chief), Bishop, M.P. (Ed.), Treatise on Geomorphology. Academic Press, San Diego, CA, vol. 3, Remote Sensing and GIScience in Geomorphology, pp. 259–298, 2013.

Yamazaki, D., O'Loughlin, F., Trigg, M.A., Miller, Z.F., Pavelsky, T.M. & Bates, P.D. 2014, "Development of the global width database for large rivers", Water Resour. Res., 50, 3467–3480, doi: 10.1002/2013WR014664, 2014.

Yitzhaki, S., Schechtman, E. The Gini Methodology. A Primer on a Statistical Methodology. 2013. Springer Series in Statistics. Volume 272, 2013, ISBN: 978-1-4614-4720-7.

Zaraj, Z., Zambrano-Bigiarini, M., Salamon, P. Burek, P., Gentile, A., and Bianchi, A.: Calibration of the LISFLOOD hydrological model for Europe. Calibration Round 2013JRC Technical Report, European Commission, Joint Research Centre. Ispra, Italy, 2013**.**

Zhang, Y., Hong, Y., Wang, X., Gourley, J.J., Gao, J., Vergara, H.J. and Yong, B.: Assimilation of passive microwave streamflow signals for improving flood forecasting: A first study in Cubango River Basin, Africa. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 6(6), pp. 2375-2390, 2013.

Zhu, Z., Bi, J., Pan, Y., Ganguly, S., Anav, A., Xu, L., Samanta, A., Piao, S., Nemani, R.R. & Myneni, R.B.: Global data sets of vegetation leaf area index (LAI)3g and fraction of photosynthetically active radiation (FPAR)3g derived from global inventory modeling and mapping studies (GIMMS) normalized difference vegetation index (NDVI3G) for the period 1981 to 2011, Remote Sensing, vol. 5, no. 2, pp. 927-948, 2013.

Table 1. Number of catchments by continent and range of upstream areas for the located stations.[1]Stations used for calibration and validation.[2] South African upstream catchment areas are not available.

Continent	**Number of satellite locations for extraction (n=398)**	**Number of stations for calibration (n=322)**	**Number of Catchment[1]**	**Upstream catchment areas (km^2) Approx. range**
Africa	75	51	21	46990 – 850500[2]
Asia	23	3	4	7150 - 11000
Europe	13	7	3	9000 - 132000
North America	207	183	86	5300 - 1850000
South America	80	78	38	1400 - 4680000

Table 2. Climate and land cover type of the 322 sites selected for the calibration and validation, aggregated by continent, climate, and land cover. [1] Vegetation means a combination of grassland, shrubland and forest. [2]Types of land cover and climate where the number of locations in each type was very low (e.g. 3) were excluded for their respective variables analysis as they will not be representative on a global scale.

Climate	**Africa**	**Asia**	**Europe**	**North America**	**South America**	**Total**
Arid	30			25		55
Tropical	10				75	85
Temperate	11		3	51	3	68
Cold		3	4	104		111
Polar[2]				3		3
Total	51	3	7	183	78	322
Land cover	**Africa**	**Asia**	**Europe**	**North America**	**South America**	**Total**
Open Forest	4			23		27
Closed to Open Forest	16	1	1	16	41	75
Closed Forest				33		33
Mosaic Vegetation predominant [1]	19	2		47	24	92
Mosaic cropland or grassland predominant	5		1	26	9	41
Rainfed crop			4	5	4	13
Sparse vegetation	2			14		16
Sparse vegetation+crops	5			8		13
Urban			1	10		11
Bare areas[2]				1		1
Total	51	3	7	183	78	322

Table A 1. Studied land cover types from GlobCover (2009) aggregated into broader categorical classes by type and vegetation density.

Label	Aggregated classes
Rainfed croplands	Rainfed croplands
Sparse (<15%) vegetation	Sparse vegetation
Closed to open (>15%) broadleaved evergreen or semi-deciduous forest (>5m)	Closed to open forest
Closed to open (>15%) mixed broadleaved and needleleaved forest (>5m)	Closed to open forest
Closed to open (>15%) (broadleaved or needleleaved, evergreen or deciduous) shrubland (<5m)	Closed to open forest
Closed to open (>15%) herbaceous vegetation (grassland, savannahs or lichens/mosses)	Closed to open forest
Closed to open (>15%) broadleaved forest regularly flooded (semi-permanently or temporarily) - Fresh or brackish water	Closed to open forest
Closed to open (>15%) grassland or woody vegetation on regularly flooded or waterlogged soil - Fresh, brackish or saline water	Closed to open forest
Open (15-40%) broadleaved deciduous forest/woodland (>5m)	Open forest
Open (15-40%) needleleaved deciduous or evergreen forest (>5m)	Open forest
Mosaic cropland (50-70%) / vegetation (grassland/shrubland/forest) (20-50%)	Mosaic cropland or grassland
Mosaic grassland (50-70%) / forest or shrubland (20-50%)	Mosaic cropland or grassland
Mosaic vegetation (grassland/shrubland/forest) (50-70%) / cropland (20-50%)	Mosaic vegetation predominant
Mosaic forest or shrubland (50-70%) / grassland (20-50%)	Mosaic vegetation predominant
Closed (>40%) broadleaved deciduous forest (>5m)	Closed forest
Closed (>40%) needleleaved evergreen forest (>5m)	Closed forest
Closed (>40%) broadleaved forest or shrubland permanently flooded - Saline or brackish water	Closed forest
Artificial surfaces and associated areas (Urban areas >50%)	Urban

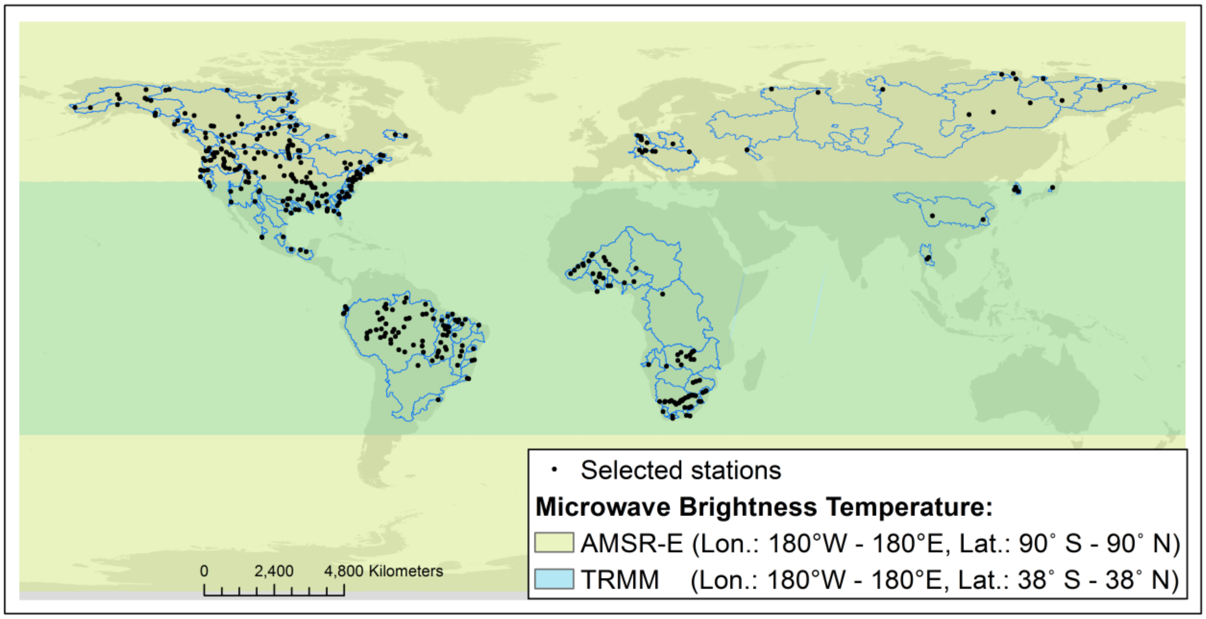
• Selected stations
Microwave Brightness Temperature:
AMSR-E (Lon.: 180°W - 180°E, Lat.: 90° S - 90° N)
TRMM (Lon.: 180°W - 180°E, Lat.: 38° S - 38° N)
0
2,400
4,800 Kilometers

Figure 1. Location of selected stations (398) and corresponding river basins (109). TRMM and AMSR-E brightness temperature product extents are also provided.

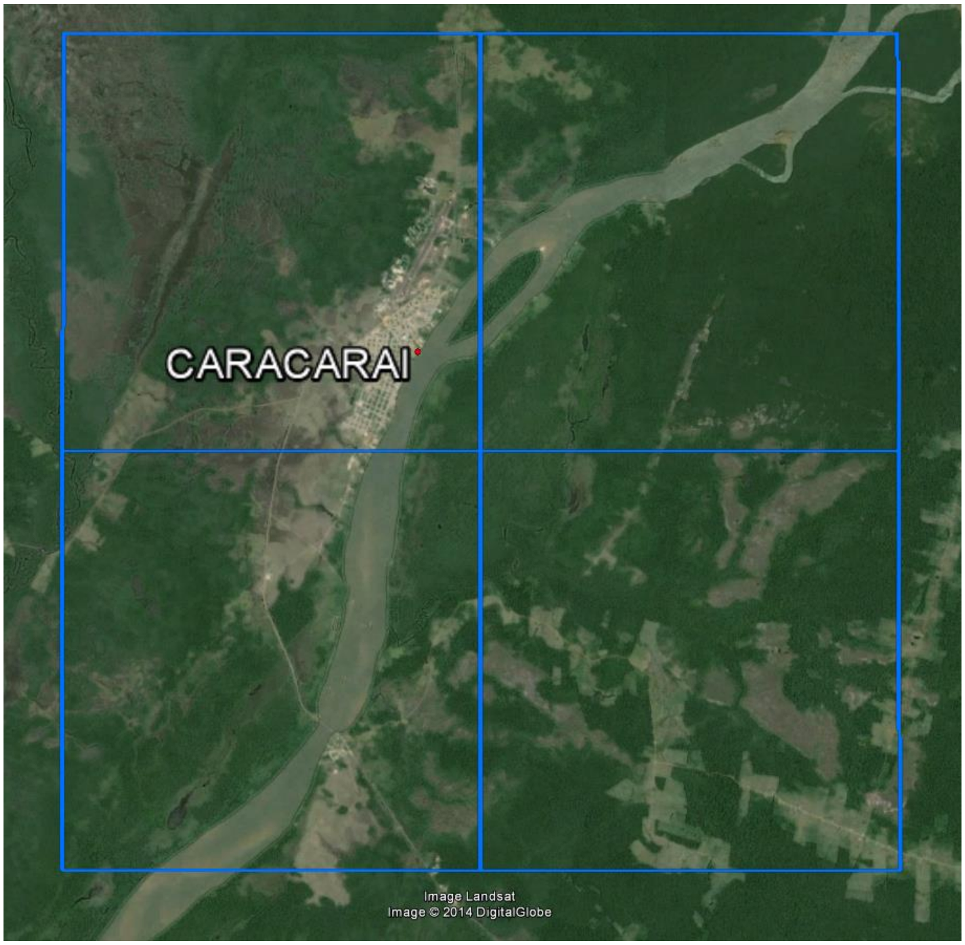

Figure 2. Example of a measurement site: Caracarai station (Rio Branco Catchment, Brazil). The blue rectangles outline the measurement pixels and background image is from 2014 Google (Landsat, DigitalGlobe).

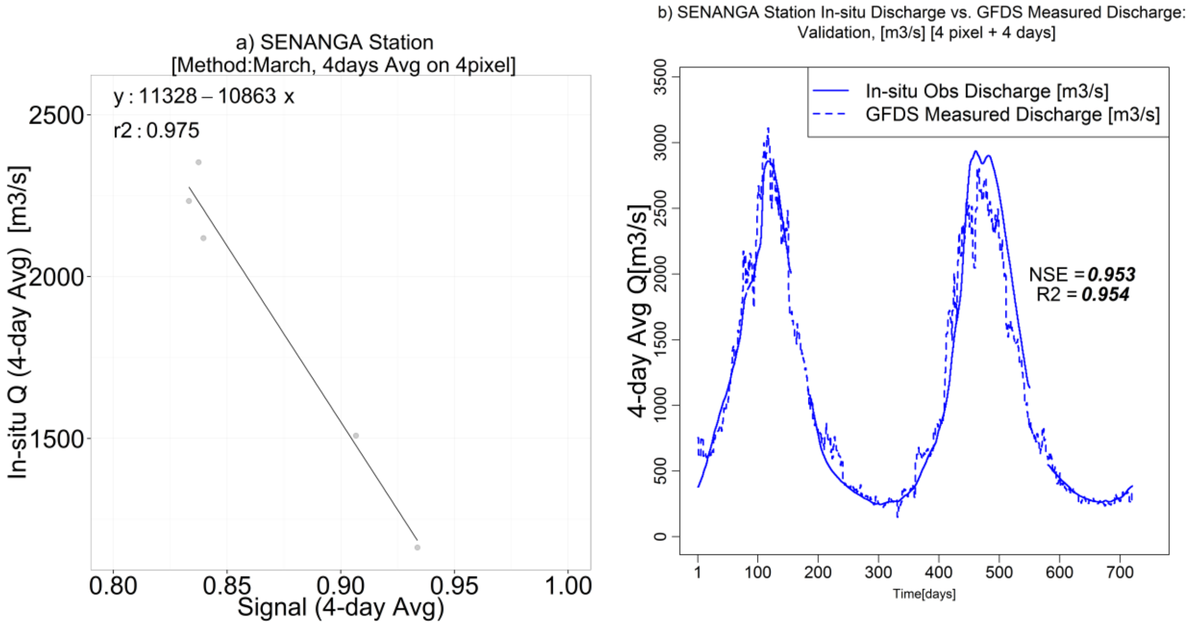

Figure 3. a) scatterplot for the Senanga station (Long 23.25, Lat. -16.116) in the Zambezi River (Africa). Monthly mean for March from 1998 up to 2002. b) Validation hydrograph for 2003-2004 and skill scores for Senanga. The (monthly) rating equations were used to calibrate the signal into discharge units. Different rating equations were used for different months.

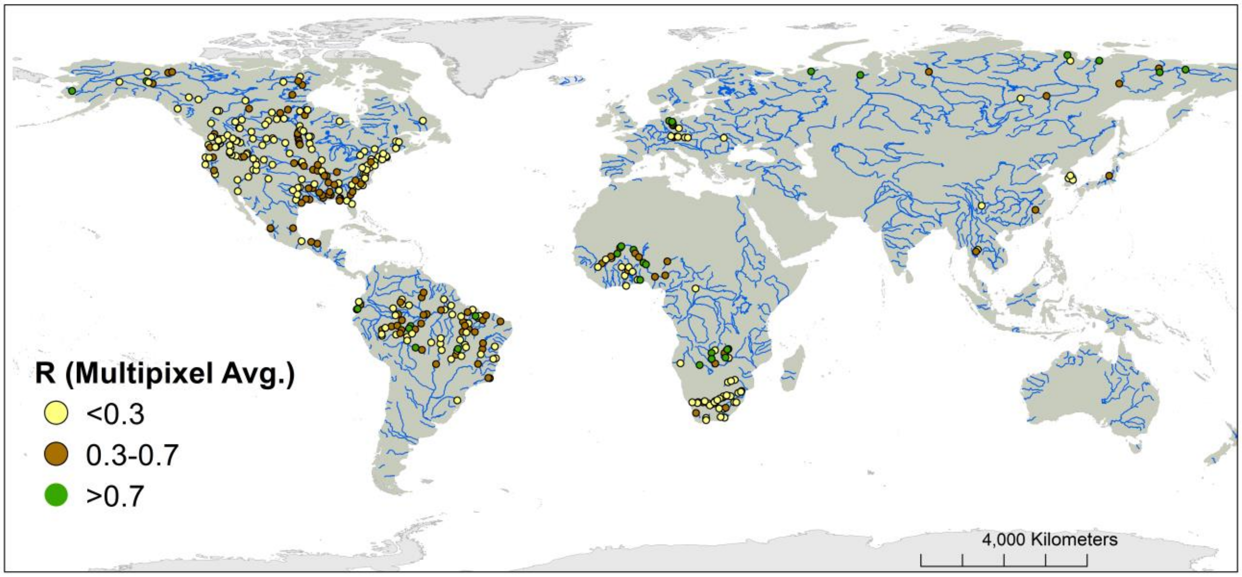

Figure 4. Location of stations and R skill score between in situ observed discharge and satellite signal (4 days and 4 pixels average). Globally, 169 sites have R>0.3, of which 42 have R>0.5.

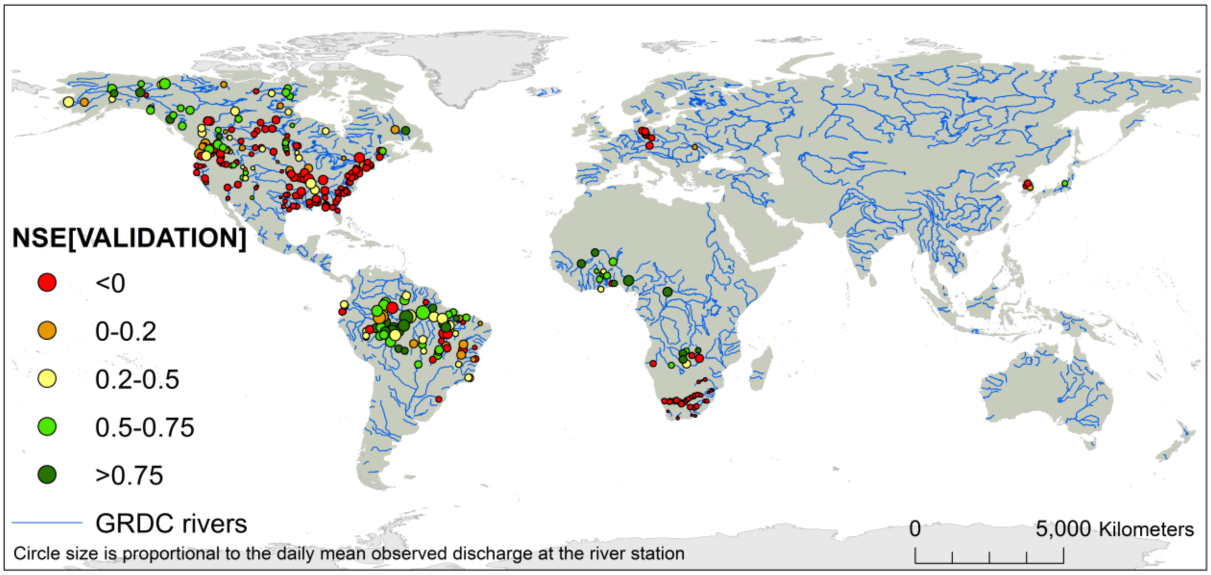
NSE[VALIDATION]
<0
0-0.2
0.2-0.5
0.5-0.75
>0.75
GRDC rivers
Circle size is proportional to the daily mean observed discharge at the river station
0
5,000 Kilometers

Figure 5. Nash-Sutcliffe efficiency of the validation (n= 332 stations). Globally, 154 stations have NSE>0 of which 80 stations have NSE> 0.50.

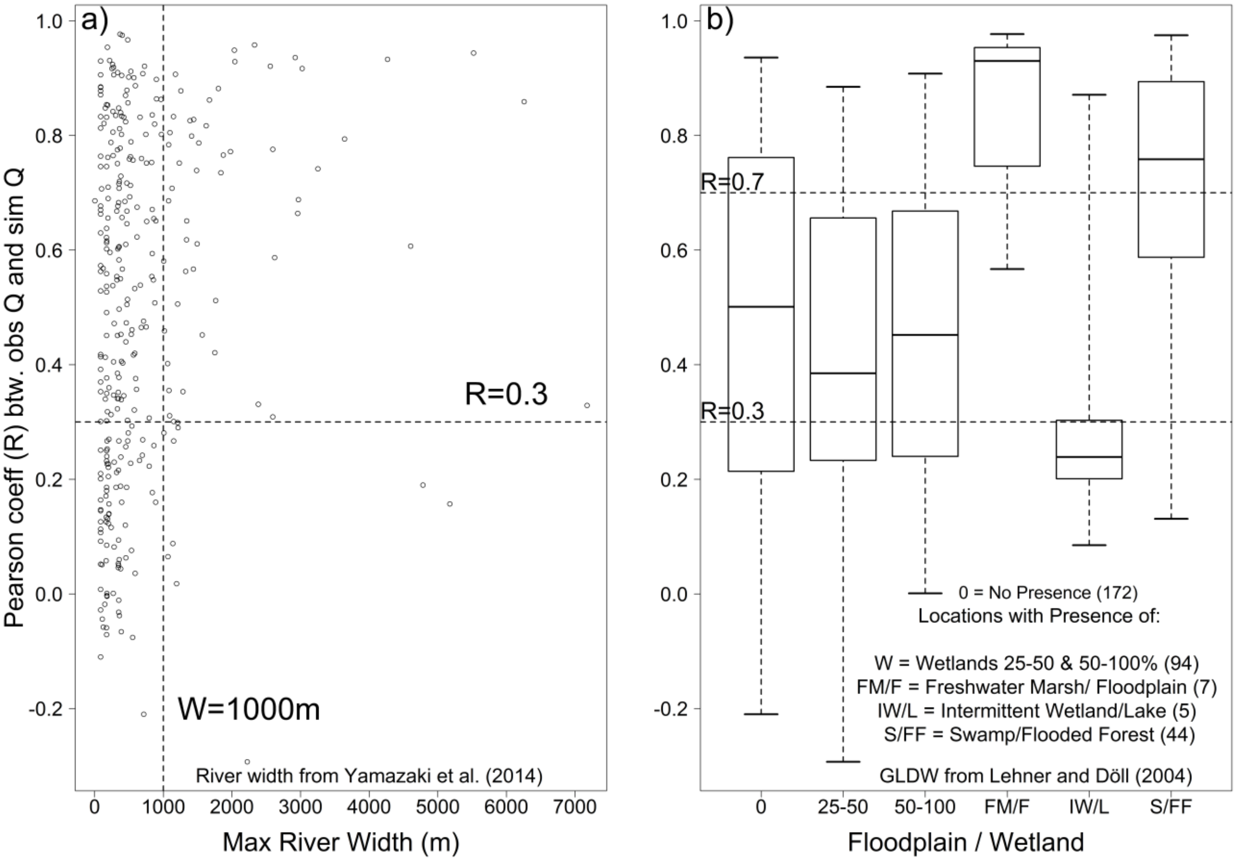

Figure 6. a) relationship between R obtained from the validation of satellite measured discharge and the maximum river width for each location; b) relationship between the same R score and the presence of significant floodplains, flooded forest and wetlands Horizontal dotted line shows the R=0.3 and R=0.7 threshold, the vertical line is the river width equal to 1km.

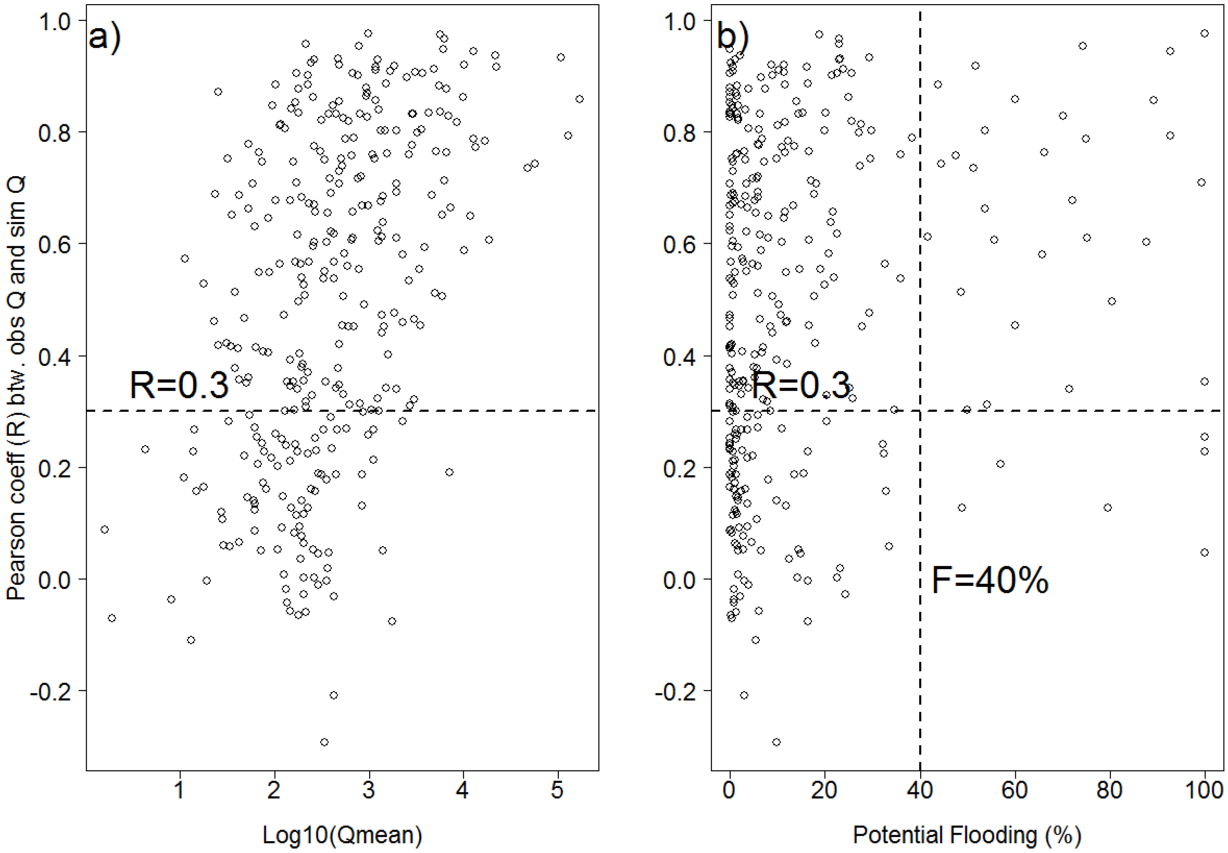

Figure 7. a) relationship between R obtained from the validation of satellite measured discharge and the mean in situ observed discharge (log10 displayed) for each station; b) relationship between the same R score and the potential percentage of flooded area per pixel for a 100 year return period flood event (Pappenberger et al., 2012). Horizontal dotted line shows the R=0.3 threshold, the vertical line is the 40% potential flooding threshold.

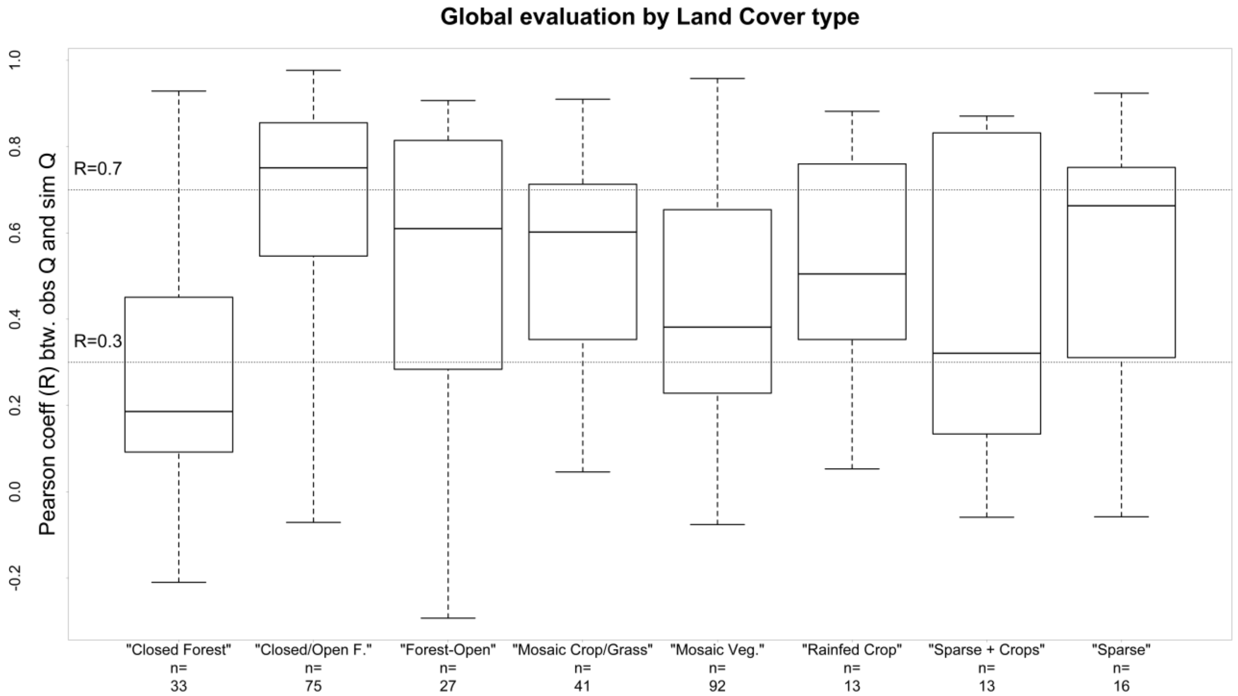

Figure 8. Global evaluation of the R score obtained during the validation and its classification by the land cover type of the stations. Land cover type were aggregated from the GlobCover (2009) and modified by visual check with Google maps. Note that artificial and bare land cover were excluded on this figure.

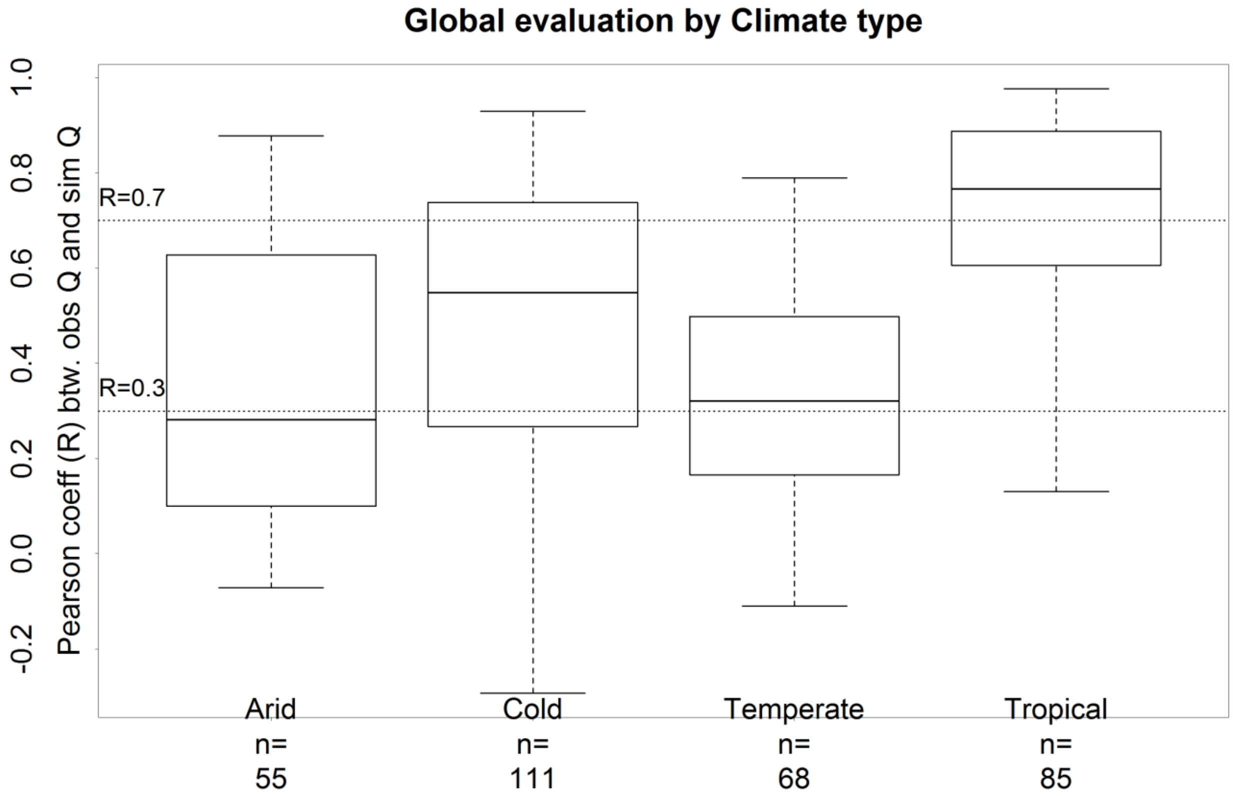

Figure 9. Global evaluation of the R score obtained during the validation and its classification – only main types-by the Köppen-Geiger climate area (Peel et al. 2007). Note that polar climate was excluded from this analysis as only three stations felt into this category.

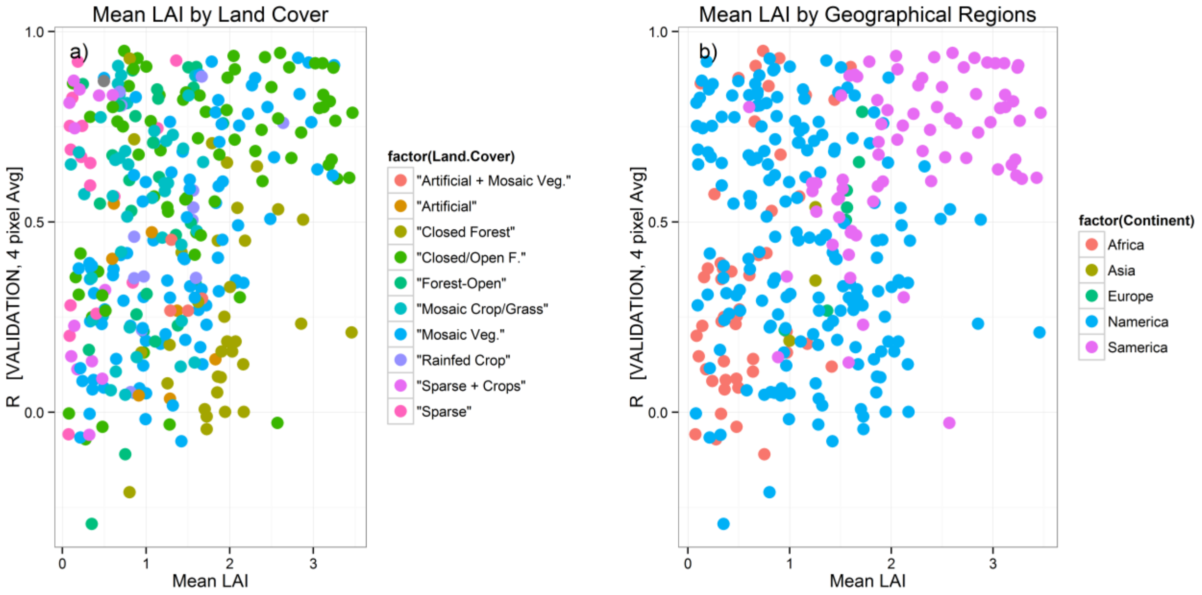

Figure 10. Evaluation of the R score obtained during the validation and its classification by Leaf Area Index (LAI), also a factor of land cover and geographical regions.

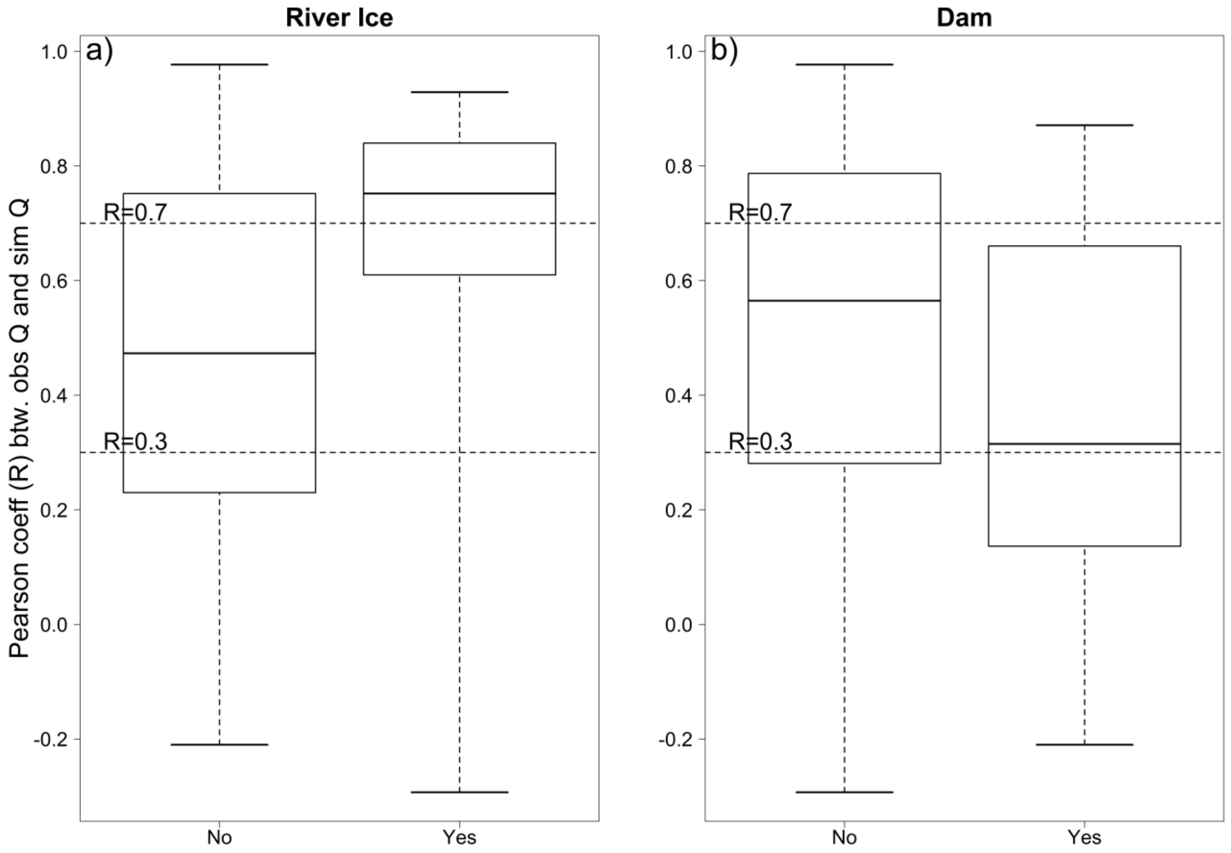

Figure 11. Evaluation of the R score obtained during the validation and its classification by a) presence or not of a river ice (Brown et al., 2002), b) presence or absence of a nearby dam or hydraulic control infrastructure using the Global Reservoir and Dam (GRanD) (Lehner et al., 2008)and visual check from Google maps.. To note that for the validated locations, all stations with river ice and most of them with dams and are located in North America.

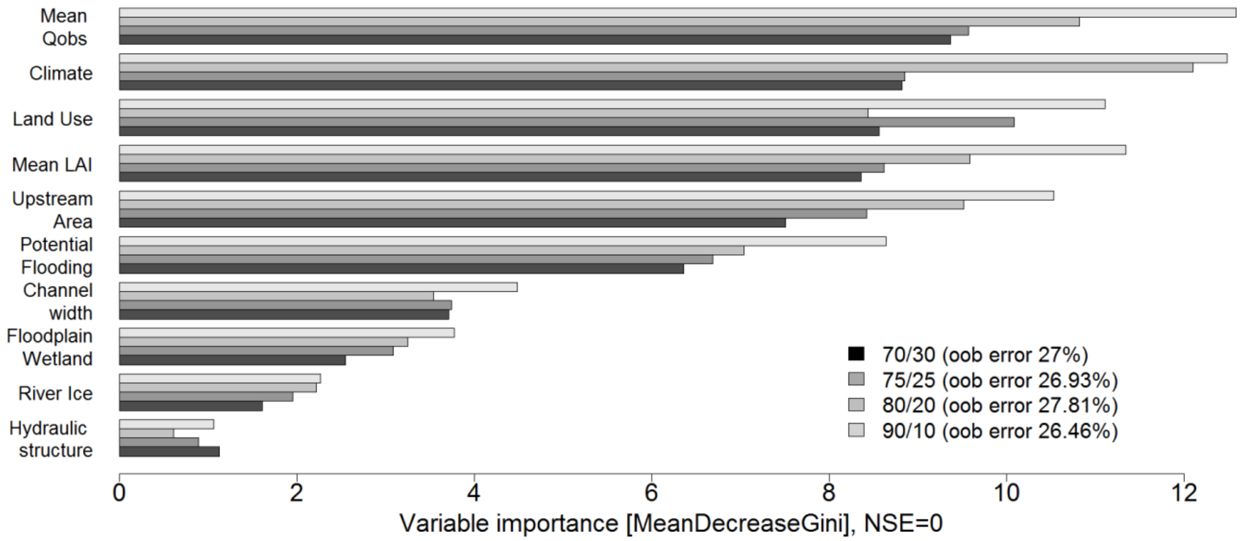

Figure 12. Average variable importance of 200 runs using the Random Forest methodology. Nash-Sutcliffe score was chosen as a quality index to categorised the stations as true (good predictive) or the stations as false (poor predictive). With a threshold of NSE=0, we have about 50% of the stations above and below that value. Results are shown for the different training and test groups. For all the test groups and runs, the average highest variable importance was obtained for mean observed discharge, climatic region, land cover/ mean LAI and upstream cacthment area, and the lowest for dam/hydraulic structure presence and river ice.