Eberhard Karls Universität Tübingen
Mathematisch-Naturwissenschaftliche Fakultät
Wilhelm-Schickard-Institut für Informatik

Bachelor Thesis Mediainformatics

ENSO impact on river discharge in South America

Markus Deppner

8th September 2021

Reviewer

Dr. Bedartha Goswami
Machine Learning in Climate Science
Cluster of Excellence "Machine Learning"
University of Tübingen

Markus Deppner
ENSO impact on river discharge in South America
Bachelor Thesis Mediainformatics
Eberhard Karls Universität Tübingen
Thesis period: from 17th May 2021 until 13th September 2021

Abstract

The impact of the El Niño Southern Oscillation on hydrological systems such as rivers is well known, but most existing studies are severely limited by data coverage. Time series of gauging stations fade in and out over time, which makes hydrological large scale and long time analysis challenging. Especially when investigating rarely occurring extreme events missing data is either constraining the spatial area or the timespan of the study. A data driven analysis of the ENSO impact on river discharge in South America is therefore missing. The purpose of this study is to overcome this spatio-temporal trade-off and fill the gap of such an assessment using in situ data. We use Gaussian Process Regression to infer missing streamflow data based on temporal correlations of stations with missing values to others with data. By using 216 stations, from the "Global Streamflow Indices and Metadata Archive", that initially cover the 56 year timespan, we were able to extend the data by over 11-fold as we could estimate missing data for 2210 stations. The spatial impact of strong ENSO events between 1960 and 2016 has then been analysed with the extended dataset. For both La Niña and El Niño events the area of their spatial impact was much larger than its impact on climate would suggest. Eastern Pacific El Niños had an exceptionally strong impact, like the strong event of 1982/83, which was by far the most intense event in our study. The top-level maps reveal that strong ENSO events are capable of impacting peak river discharge and floods over the whole continent whereas the individual impact maps show very different characteristics and severity in its impact pattern.

1 Introduction

Floods are among the most devastating natural hazards, claiming multiple hundreds of lives per year and causing enormous economical and ecological damage in the affected regions (e.g. Ritchie and Roser, 2014; Jonkman, 2005; Dodangeh et al., 2020). It often takes multiple years or decades to fully restore the damages caused by the flood. Especially developing countries are more exposed to the effects of floods, recording the highest number of deaths and struggle even more to overcome the long term damages of floods (Zorn, 2018). But even developed countries are not fully prepared to deal with the devastation of floods, despite early predictions and warning systems, as demonstrated by the recent floods in July 2021 in Belgium and Germany. Over the years, societies have developed mitigation measures such as physical barriers like dams or complex warning systems and forecasting models to minimize the vulnerability and the effect of floods on societies (e.g. Schanze et al., 2006; Hirabayashi et al., 2013). Enhancing flood forecasting, risk assessment, and warning models is a main motivation for hydrological researchers around the globe to reduce humanitarian and socioeconomic damage. Especially with regard to an increase in flood frequency, intensity and an increase in global flood risk in the next decades under a changing and warming climate (e.g. Hirabayashi et al., 2013; Alfieri et al., 2017), this becomes even more important. However flood forecasting is subject to large uncertainties due to the intrinsic uncertainty of meteorological forecasts and of structural parameters of rainfall-runoff models (Dietrich et al., 2009). Therefore, past floods, their driving mechanisms and their courses are constantly studied to reduce this uncertainty by gaining a better understanding of the processes and enhance flood forecasting and risk assessment models. Guimarães Nobre et al. (2019) claim that the forecasts of natural hazards such as floods or droughts can particularly benefit from a better understanding of the ENSO and its embedding into climate models. However hydrological analyses and studies of past floods with in situ data are often challenging, as a lot of studies are severely limited due to insufficient data availability and missing values. This can lead to floods not being represented in the data, because gauging stations might not have been actively recording during the time of occurrence. This is the reason why flooding events cannot be reasonably analysed using in situ data alone. The problem of missing data originates from a scattered installation of gauging stations at different times in the past and an inconsequent maintenance which is why many stations do not provide continuous measurement data. This results in time series of stations fading in and out over time making large-scale hydrological analyses such a challenging task. Due to the constant trade-off between temporal and spatial coverage, hydrological analyses with in situ data can usually only be reasonably performed on a basin scale. To investigate larger regions or even perform analyses on a global scale, the outputs of climate simulations are typically

used as primary data source (e.g. Yamazaki et al., 2018; van Vliet et al., 2013; Nohara et al., 2006). However, to assess the magnitude of individual drivers on river discharge, in situ data is preferred as it exhibits less uncertainty than modeled data. This study performs such a hydrological data-driven analysis for the entire continent of South America, and aims to shed light on the question of how and where the El Niño Southern Oscillation (ENSO) impacts peak river discharge. To the best of our knowledge, such a large-scale analysis using in-situ data has not been undertaken yet. This is why we were eager to fill this gap and assess ENSO's impact on rivers in South America.

The ENSO is one of the most prominent patterns for inter-annual climate variability around the globe (Emerton et al., 2017) and has a climatological and socioeconomic impact worldwide (McPhaden et al., 2006). It impacts river discharge in a direct and indirect manner as the shift in atmospheric circulation that comes along with the ENSO leads to sea surface temperature (SST) anomalies in the Pacific and induces precipitation and temperature anomalies around the globe. This is a relation quite well studied and increasingly well understood (Dilley and Heyman, 1995). In the year 1995, Dilley and Heyman already established a connection between disasters such as droughts or floods and the ENSO, especially in those regions that are affected by ENSO-induced temperature and precipitation anomalies. While this holds for many basins in South America, Yan et al. (2020) showed that this precipitation-flood relation does not hold for all basins affected by ENSO-driven precipitation anomalies and states that flood indices such as flood frequency or flood duration tend to be correlated stronger to the ENSO than precipitation. Also, the nonlinearity between precipitation and flood hazards has been shown as river discharge is a complex system, whose reduction to precipitation anomalies would be a strong simplification (Stephens et al., 2015). Streamflow is an integral part of the interplay in the hydrological water cycle, where each component impacts the other. That is one reason why modelling river discharge is such a challenging task as this interplay varies from region to region and from river to river. Also natural factors such as the drainage basin, antecedent rainfall and moisture, the type of soil and rock, the amount of precipitation, temperature and vegetation are all drivers of runoff and therefore drivers of river discharge (e.g. Yang et al., 2019; USGS.gov, 2021). In addition to the natural causes, anthropogenic factors can drastically affect runoff and river discharge as well, for example through urbanisation, the construction of dams, change of land use or deforestation among many others.
As both, the ENSO as well as river discharge or floods are multilayered and complex phenomena we aim to investigate not individual features, but the overall impact of ENSO on river discharge. We present a top level picture on the regions where ENSO impacts peak river discharge and is capable to induce floods in South America, which might possibly enhance the representation of

ENSO in flood risk assessment and flood forecasting models (Guimarães Nobre et al., 2019). This study focuses on the region of South America where gauging stations show high temporal variability, ranging from one to 116 years of data coverage. However, investigations on ENSO's impact require a long temporal coverage as it has a naturally low frequency of two to seven years and strong ENSO events occur even less frequent. Data has been reconstructed and inferred for stations that were no longer maintained or were installed only in recent years. By extending the timespan of individual stations, we were able to undertake such a large-scale data driven approach. We train a complex function for each target time series with missing values, based on its highest correlated time series that contain data over the whole timespan. Technically, this is a classic regression task for supervised learning as we desire a function to a given input-output mapping. Determining the characteristics of such a function a priori is unreasonable, which is one of the main motivations Gaussian Process Regression (GPR) has been used for this regression task. Gaussian Processes (GPs) can be understood as a distribution over functions (Rasmussen and Williams, 2006), which is why all possible complex and simple functions are respected and taken into account. Also, GPR provides a measure of uncertainty for each inferred value in our target time series, which is another main advantage to other classical regression tasks. The work of Sun et al. (2014) compared the performance of forecasting streamflow data with GPR to linear regression and artificial neuronal network models and GPR outperformed both models in most cases, which fortified the decision in using GPR for our study. By extending our data coverage through interpolating missing values with GPs we attempt to present a possible solution to deal with the spatio-temporal trade-off and the problem of missing data when working in a hydrological context. The main motivation for this study is to examine the ENSO impact on South America with regard to peak river discharge. Also, we aim to contribute to the disentanglement of the interplay of climatological driving forces and their magnitude on river discharge by assessing the impact on river discharge during ENSO phases. Getting a better understanding about where the ENSO impacts floods and to what extent can possibly enhance flood forecasting and flood risk assessment tools, as ENSO's impact as a driving force can be better understood and embedded in models.

2 Data and Methods

2.1 Global Streamflow Indices and Metadata Archive

The "Global Streamflow Indices and Metadata Archive" (GSIM) published in April 2018 by Lukas Gudmundsson, Hong Xuan Do, Michael Leonard, and Seth Westra is a hydrological streamflow dataset on a global scale, containing streamflow data from in situ gauging stations around the world. With the release of this dataset, a large step towards dismantling barriers of the accessibility of hydrological data has been made, pushing transparent research and open accessibility forward at the same time. The streamflow data from various institutes has been cleaned, standardized and assessed by uniform quality measures to create this dataset. Merging scattered datasets around the world to create one central database and providing this archive for free makes the GSIM the first of its kind and is a big contribution to the hydrological community. The highest resolution in the GSIM is monthly, which is the resolution used for this study. As some organisations prohibited the publication of daily data, monthly, seasonal and yearly datasets have been derived from daily streamflow measurements. The GSIM is composed of 30,959 streamflow gauging stations from which 3,449 are located in South America. The spatial distribution of stations with missing values during the timespan from January 1960 until end of May 2016 and the ones with continuous data is presented in Figure 1.

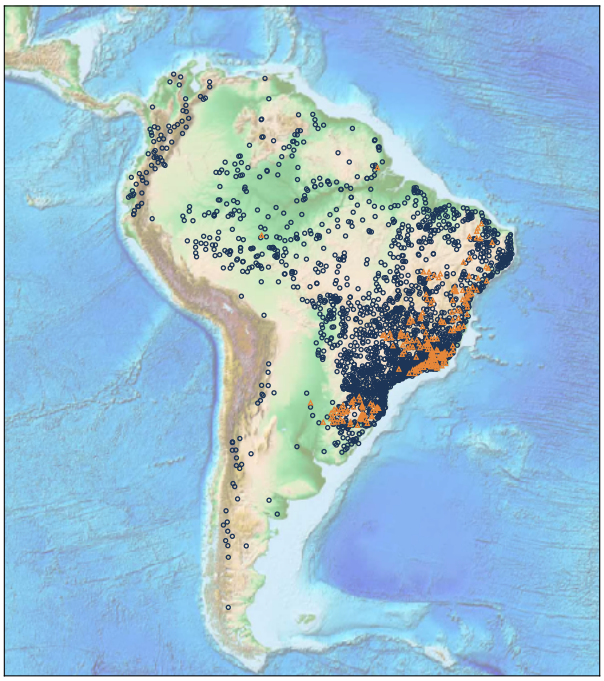


Figure 1: Spatial distribution of stations initially covering the full timespan of 56 years from 1960 until 2016 (216 stations marked in orange) and stations that contained missing values during this time (2210 stations marked in blue).

The average temporal coverage of a GSIM station in South America amounts to 29.3 years, which is the second lowest continental average and more than 23% below the overall average temporal coverage (Do et al., 2018). Stations in South America show a large temporal variety from one year up to 116 years. This originates from a scattered and inhomogeneous installation of gauging stations in combination with an insufficient maintenance. With this work we strongly rely on a large temporal coverage such that we are able to include a sufficient number of events in the analysis.

The Oceanic Niño Index (ONI) in combination with the standard definition, where at least three consecutive months of ± 0.5 in the Niño 3.4 region are required to be declared as an El Niño or a La Niña event (NOOA's Climate Prediction Center, 2021), has been used to indicate the strongest ENSO events during our timespan of 56 years. The determined years are mainly consistent with other indices such as the Southern Oscillation Index or the Multivariate ENSO Index (MEI) (McPhaden, 2020). The years and their respective ONI are listed in Table 1.

El Niño	ONI	La Niña	ONI
1965/66	2.0	1973/74	-2.0
1972/73	2.1	1974/76	-1.7
1982/83	2.2	1988/89	-1.8
1997/98	2.4	1998/01	-1.7
2015/16	2.6	2010/11	-1.6

Table 1: The five strongest El Niño and La Niña events and their absolute maximum values achieved by the ONI during the active phase. An event has been declared due to the standard definition of 3 consecutive months of +/- 0.5 ERSST.v5 SST anomalies in the Niño 3.4 region (NOOA's Climate Prediction Center, 2021).

2.2 Spearman's Rank Correlation

Based on 216 stations that initially covered the 56 year timespan, missing values for 2,210 stations were inferred that show a minimum overlap of 60 months of valid data with our timespan. For each target timeseries the Spearman's Rank Correlation (SRC) has been computed on the temporal overlap with stations covering the whole timespan. The ten highest correlated timeseries were used to train the Gaussian Process model. We chose SRC to compute the correlation between streamflow timeseries as it is more robust to outliers than regular Pearson's Correlation by its transformation of the values of a timeseries into their ranks . Also, it shows higher correlation values for non-linear but monotonic correlated timeseries, which is more robust when analysing timeseries of natural phenomena with large variability as they are not necessarily linearly related.

The ten most correlated timeseries were included in the training process for the GP model. Test runs to verify the amount of stations to include in the model were made, ranging from one to 200 stations. The distribution of the respective root mean squared errors (RMSE) were analysed and in general fewer timeseries achieved lower RMSE values and better results. The number of ten was reasonable and passed the robustness check of $\pm$ 50 % where the distribution of RMSE did barely vary which is shown in Figure S2.

2.3 Gaussian Normalisation

Streamflow data is not normally distributed and shows a large variety among stations in terms of absolute discharge. That is why raw streamflow data is challenging to handle for the GP. To normalize the data, a transformation into percentiles has been performed on each timeseries before transferring into a Gaussian distribution. This transformation preserves the relative dynamics, but discards the absolute discharge values of a time series. This comes along with convenient Gaussian properties which enables us to train the GP model with ease. Figure 2 shows an exemplary excerpt of a timeseries during each transformation step and its corresponding distribution.

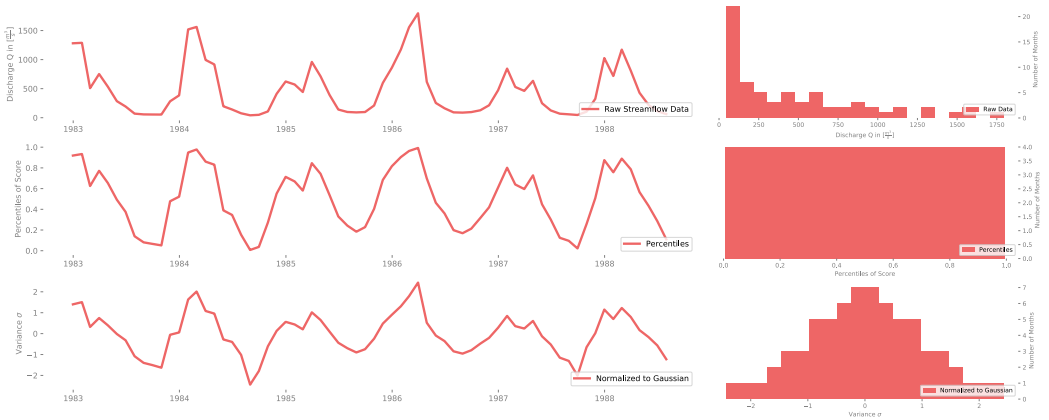


Figure 2: Transformation of a time series during each transformation step from measured river discharge $[\frac{m^3}{s}]$ into percentiles of score and finally into a zero mean Gaussian distribution (left), and its corresponding distribution in form of a histogram (right).

2.4 Gaussian Process Regression

The derivations and equations in this section are cited and summarized from Rasmussen and Williams (2006) and Sun et al. (2014).

Gaussian Processes (GPs) are a probabilistic machine learning tool to learn input-output mappings from empirical data in a supervised learning fashion (Rasmussen and Williams, 2006). GPs can be used as a tool for classification

for discrete outputs and also for regression when continuous outputs are required, as they are in our case. A classical linear regression task aims to learn a function f such that it explains the relationship between an input variable $x \in \mathbb{R}^d$ and its target variable y which is mostly a scalar

$$y = f(x) + \epsilon. \tag{1}$$

The variable ϵ is an error term which should be kept as small as possible for the best possible performance. The function f can be expressed as a combination of a set of M basis functions $\phi_j(x)_{j=1}^M$ which can be linear or non-linear and a set of weights $w = [w_1, ..., w_M]^T$, where each basis function is scaled by its corresponding weight

$$y = \sum_{j=1}^{M} w_j \phi_j(x) + \epsilon. \tag{2}$$

The weights are trained such that they explain the input output mapping (x,y) best. For a linear regression, two parameters would be trained (describing the slope and the intersection), for a quadratic function three parameters. A main advantage of GPR is that a prior limitation to a set of parameters or functions is not needed as GPs are non-parametric models. This means that an infinite amount of parameters can be trained, which includes all possible functions that fit the data. For modelling streamflow time series a prior constraint on the characteristics of the function is not reasonable which is why GPR is well suited for our regression task. In order to train the GP model to the whole dataset $X = \{x_i\}_{i=1}^N$ which is composed of N observations and the respective target vector $Y = [y_1, ..., y_N]$ such that it fits the whole dataset, $f = \{\hat{f}(x_i, w)\}_{i=1}^N$ now defines the model outputs for our input dataset X for

$$\hat{f}(x_i, w) = \sum_{j=1}^{M} w_j \phi_j(x_i) \qquad i = 1, ..., N. \tag{3}$$

We can rewrite Equation 3 by using an N x M design matrix Φ which holds the evaluation all M basis functions for its respective input x_i in each row. A row of Φ is therefore written as $\phi_j = [\phi_1(x_i), \phi_2(x_i), ..., \phi_M(x_i)] \quad j = 1, ..., N$ and the whole model output as

$$f = \Phi w. \tag{4}$$

The design matrix Φ is also used for defining the covariance matrix $K \in \mathbb{R}^{NxN}$. The symmetric and positive definite covariance matrix K is defined as

$$K = \Phi E(ww^T)\Phi^T = \Phi \Sigma_w \Phi^T, \tag{5}$$

with Σ_w as the covariance matrix of the weight vector w.

On a top level Gaussian Processes can be understood as a distribution over functions that is completely specified by its second-order statistics, its mean $m(x)$ and its covariance function $k(x, x')$ (Rasmussen and Williams, 2006)

$$f(x) \sim GP(m(x), k(x, x')). \tag{6}$$

As the GP is defined to be a collection of finite sets of random variables which follow Gaussian distribution, the prior of f is Gaussian distributed. The prior is defined by its covariance Matrix K and its hyperparameters of the covariance function are denoted as θ. For simplicity a mean of zero is chosen here

$$p(f \mid X, \theta) \sim \mathcal{N}(0, K). \tag{7}$$

For a Gaussian distributed error term ϵ the vector of our target variables y is also Gaussian

$$p(y \mid f, \sigma^2) \sim \mathcal{N}(f, \sigma^2 I), \tag{8}$$

with I as the identity matrix and σ^2 as the variance of the error.
The desired posterior distribution after applying Bayes' rule is defined as:

$$p(f \mid y, X, \theta, \sigma^2) = \frac{p(y \mid f, \sigma^2)\, p(f \mid X, \theta)}{p(y \mid X, \theta, \sigma^2)}. \tag{9}$$

Just like the prior and the likelihood the posterior also follows a Gaussian distribution. After substituting the prior and the likelihood into the posterior, the following closed formulas for mean and covariance are derived:

$$\mu = K^T (K + \sigma^2 I)^{-1} y \tag{10}$$

$$\Sigma = K - K^T (K + \sigma^2 I)^{-1} K. \tag{11}$$

The mean and the covariance depend on the covariance matrix K, which is composed of Φ and its weight vector w. However, we can define K by a covariance function $k(x, x^{'})$ and thus μ and Σ do no longer depend on Φ. This shifts the focus to the definition of the covariance matrix, as we no longer need to determine individual basis functions and their weights. In theory, GPs are very sensitive to the choice of kernels, and the choice should be assessed for each application carefully as it generally determines the performance of the models prediction. In its practical application however, GPs are not very sensitive to different choices of covariance functions for time series modeling (e.g. Shi et al., 2007; Sun et al., 2014), which is consistent with the results of our kernel analysis upfront. For a subset of streamflow timeseries GP Regression models were trained with different kernels and varying dimensionality. For each kernel

the Root Mean Squared Error (RMSE) has been computed for the test set, which has shown stable and similar results for different kernels. We chose a Matérn kernel for our study as it was among the most stable kernels with low RMSE. The Matérn covariance function for two points and its distance d is given by

$$C_\nu(d) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} (\sqrt{2\nu}\frac{d}{p})^\nu K_\nu(\sqrt{2\nu}\frac{d}{p}), \tag{12}$$

where p and ν are positive parameters of the covariance and Γ is the gamma function and K_ν the modified Bessel function of the second kind (Rasmussen and Williams, 2006).

The marginal probability of the target vector y given our dataset X is obtained by integration over all possible functions f

$$p(y \mid X) = \int p(y \mid f, \sigma^2) p(f \mid X, \theta)\ df. \tag{13}$$

From this marginal probability we can compute the log marginal likelihood as

$$\log p(y \mid X) \propto -\frac{1}{2} y^T (K + \sigma^2 I)^{-1} y - \frac{1}{2} \log |K + \sigma^2 I| - \frac{N}{2} \log (2\pi). \tag{14}$$

Via a gradient-based algorithm the unknown parameters of θ and σ^2 can be derived from the log likelihood. As we computed the likelihood, the prior and the evidence we can now derive the desired posterior distribution by plugging them into Bayes theorem from Equation 9. With the posterior distribution we can now predict the Gaussian distribution of any test sample x_* conditioned on the trained model

$$p(f_* \mid x_*, y, X, \theta, \sigma^2). \tag{15}$$

The mean m and the variance ν^2 of the target Gaussian distribution can then be derived for a given test sample x_* by:

$$m(x_*) = \phi(x_*)^T \mu = k_*^T (K + \sigma^2 I)^{-1} y \tag{16}$$

$$\nu^2(x_*) = \phi(x_*^T \Sigma \phi(x_*) = k_{**} - k^T (K + \sigma^2 I)^{-1} k_*, \tag{17}$$

where $k_* = [k(x_*), x_1), ..., k(x_*, x_N]^T$ is the evaluation of the kernel-function of the new test sample in combination to every training sample we've seen so far and $k_{**} = k(x_*, x_*)$ contains the evaluation of the kernel function between the test sample and itself. Σ and μ are the mean and variance defined by the posterior distribution.

The analysis of this work was done in python and the Gaussian Processes framework GPy developed from the machine learning department at Sheffield University (GPy, 2012) was used to train and predict our GPR models. We used the predefined Matérn32 kernel for this analysis and a mean prior of zero for each model.

2.5 Similarity Analysis between ENSO events

A graphical approach to detect similarities between ENSO events, is to visually compare the maps of the spatial patterns of affected stations. This is a good approximation to assess similarity, but in order to quantify the similarity between events we determined the stations that were affected by top ten peak river discharge during the active phase of the respective ENSO event. Then the Jaccard distance has been computed on the sets of affected stations for each combination of events among the El Niño and La Niña events. The Jaccard distance is calculated by the fraction of the cut set of affected stations that show peak river discharge during both events, divided by the union of all stations affected during those two events. On the distance matrix D hierarchical clustering was performed to group similar El Niño and La Niña events together.

$$D(i,j) = J(i,j) = \frac{|I \cap J|}{|I \cup J|} \tag{18}$$

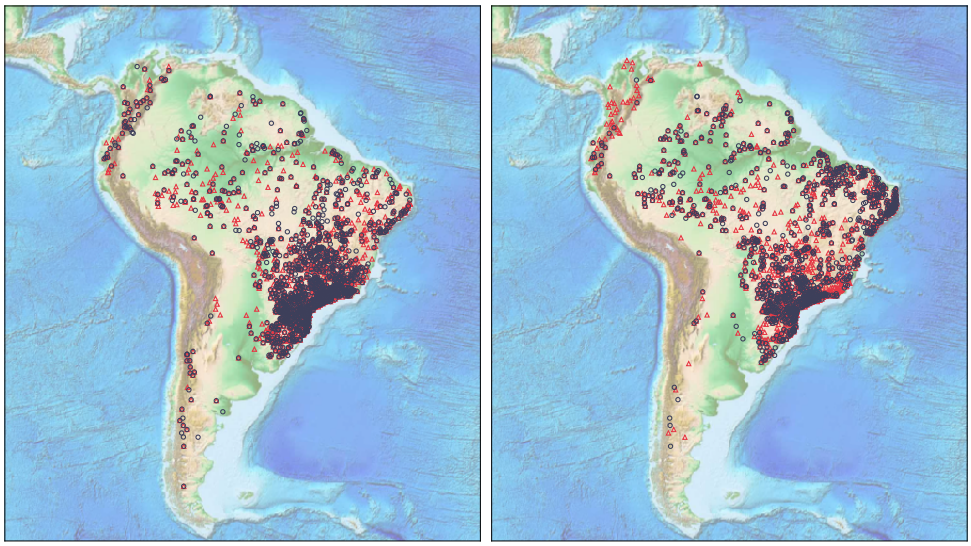
A 100-fold cross validation with 5% of removed samples in each iteration has been performed on our dataset to validate the results of the hierarchical clustering. We truncated the dendrogram such that we obtain three prominent clusters. A link was set between the events that belong to each group. This adjacency matrix enabled us to plot a weighted and undirected network graph, whose edge weights are determined by the amount of clusters in which the two events were grouped together.

2.6 Reference to code

The code for this work has been written in python, using the Gaussian Process Regression framework GPy from the University of Sheffield (GPy, 2012) and is openly accessible in this repository: `https://github.com/mdeppner/enso-streamflow-gpr`. The inferred data as well as the GSIM datasets are provided in the repository, such that the work can be reconstructed on a local machine with the jupyter notebooks.

3 Results

By interpolating data of timeseries which contain missing values, the amount of stations that can be included in this analysis was extended by over 11 fold from initially 216 to 2426 stations. This increase in data and therefore stations allowed us to undertake a data-driven large-scale hydrological analysis on streamflow and ENSO's impact on river discharge. Due to a high variability of temporal coverage of stations in the GSIM, short time series naturally show peak river discharge during their active time in which measurements were taken. As this does not necessarily represent the actual time of highest river discharge during the investigated timespan, a temporal bias would be induced when using the raw GSIM data in this study, which motivated us to infer the missing values and analyse this problem with an extended dataset.
The presentation of the results is structured from general to more specific results as we go from a top level view on the impact of El Niño and La Niña events to a more unraveled and detailed analysis of the groups of the similarity analysis. We then amplify the impact of the 1982/83 El Niño and finally present the impact of different types of ENSO events on South American rivers.



(a) Top level map of the impact of El Niño

(b) Top level map of the impact of La Niña

Figure 3: Cumulative plot of stations affected by at least one of its top ten highest monthly river discharge values measured during at least one of the five El Niños or La Niñas investigated. Stations marked with red triangles are considered to be affected based on the raw GSIM dataset and stations marked with blue circles after missing values were inferred for this timeseries.

3.1 La Niñas impact

The pattern of affected stations for which we inferred data generally aligns with the pattern of stations when using the raw GSIM dataset. Despite a number

of stations in the Colombian Andes that fall out of line and only show peak river discharge under raw data, the inferred stations generally superimpose the pattern of the raw data. Figure 3b shows the pattern of the five strongest La Niña periods combined. Stations which have at least one of its ten largest river discharge values measured during at least one of the five strongest active La Niña periods during the 56 year timespan are marked in this map. Figure S3 shows the individual pattern for each of the five strongest La Niña periods.
A large number of stations are affected by peak river discharge in the southeast of Brazil and the La Plata basin. Also a large number of stations are concentrated along the east and northeastern Atlantic coast of Brazil. Those regions are also the ones with the highest spatial concentration of stations which is depicted in the Heatmap in Figure S1. A less concentrated but still quite widely impacted area during La Niña phases is the Amazon basin where relatively many stations were affected by peak river discharge. Comparing this pattern with the regions of La Niña-induced precipitation and temperature anomalies, a clear spread farther south is noticeable (Lenssen et al., 2020). La Niñas are usually capable of inducing a wetter climate along most of Columbia, the whole Caribbean Coast, parts of north Brazil and the northern part of the Amazon region. The Argentinian Atlantic Coast and the Southern Brazil are usually drier during the active phase of a La Niña as well as a small region on the Pacific coast in central Chile (Lenssen et al., 2020). The region where La Niña is capable of inducing higher precipitation is in logical relation to where stations with peak river discharge were found. As Dilley and Heyman (1995) already stated, it seems that there exists a connection between floods and the ENSO as the regions of usually higher precipitation and temperature are also affected by peak river discharge and floods during La Niña phases. What the results show is that stations and even whole regions much farther south were affected as well. This supports the finding by Stephens et al. (2015) who stated the nonlinearity between precipitation and floodiness. It seems that La Niñas are capable of impacting river discharge and floods even in regions where it typically does not feature an increase in precipitation. This relation between floodiness or peak river discharge and precipitation will be targeted as well in the next section where the impact of El Niño is discussed.

3.2 El Niños impact

The pattern of inferred stations that show at least one of its top ten peak river discharge values during the time of at least one of the five El Niño events investigated does also align with the pattern of stations when using the raw GSIM dataset. The top level picture of El Niños impact on river discharge in South America is presented in Figure 3a which shows a cumulative plot of all stations affected during the five El Niños of interest in our study. The dense region of affected stations in the southeast of Brazil, parts of Uruguay and Paraguay is

consistent with the regions of El Niño-induced precipitation and temperature anomalies, as the region up to 20°S is usually associated with a wetter climate during an active El Niño event (Lenssen et al., 2020; Guimarães Nobre et al., 2019). The pattern of affected stations however spreads much farther north than the precipitation anomalies during an El Niño would imply. Stations in the Amazon basin were affected by peak river discharge and floods during those El Niños, although this region is usually known to have less precipitation and a drier climate during an active El Niño period (Lenssen et al., 2020). Nevertheless, our results show that strong events are capable of impacting those regions as well. The actual magnitude of El Niños impact on the Amazon basin is hard to assess with this non-uniform distribution of stationary data. That is why the general distribution needs to be taken into account when comparing spatial patterns of stations. The large widely spread impact in the Amazon basin with regard to the rather low concentration of stations in this area indicates a possibly stronger impact on this region than the scattered stations might suggest. The region with high concentration of affected stations in the southeast of Brazil on the other hand needs to be put in relation, as the general concentration of available stations in this area is very high. The south of Brazil and the northeast of Brazil are the regions with highest concentration of installed river gauges as presented in Figure S1.
The five strongest El Niño events impacted 30% more stations than the five strongest La Niña events, which might indicate that El Niño impacts South American rivers possibly stronger than La Niña. However it's hard to draw final conclusions from the absolute numbers, as the amount of stations affected during an individual event underlay high variance. Especially due to the fact that the 1982/83 El Niño had a very strong impact, which is why the absolute numbers should be considered with caution.
Figure 4 shows the individual impact pattern of every El Niño event included in our study. The South East Atlantic Coast of Brazil is affected in all of the five events and shows the most concentrated region of affected stations by peak river discharge, which is also the region of increased precipitation during an El Niño. Comparing the individual impact patterns, the magnitude of the 1982/83 event becomes clearly visible as the area of impact as well as the amount of stations is much larger than in the impact maps of the other El Niño events.

3.3 1982/83 EL Niño

During the analysis, the 1982/83 event stood out from all other strong El Niños such that it is the timespan during which the most stations were affected by peak river discharge, with both, the initial GSIM and the inferred data. With 1,040 affected inferred stations, the event of 1982/93 shows more than three times the amount of stations than any other El Niño investigated. The floods

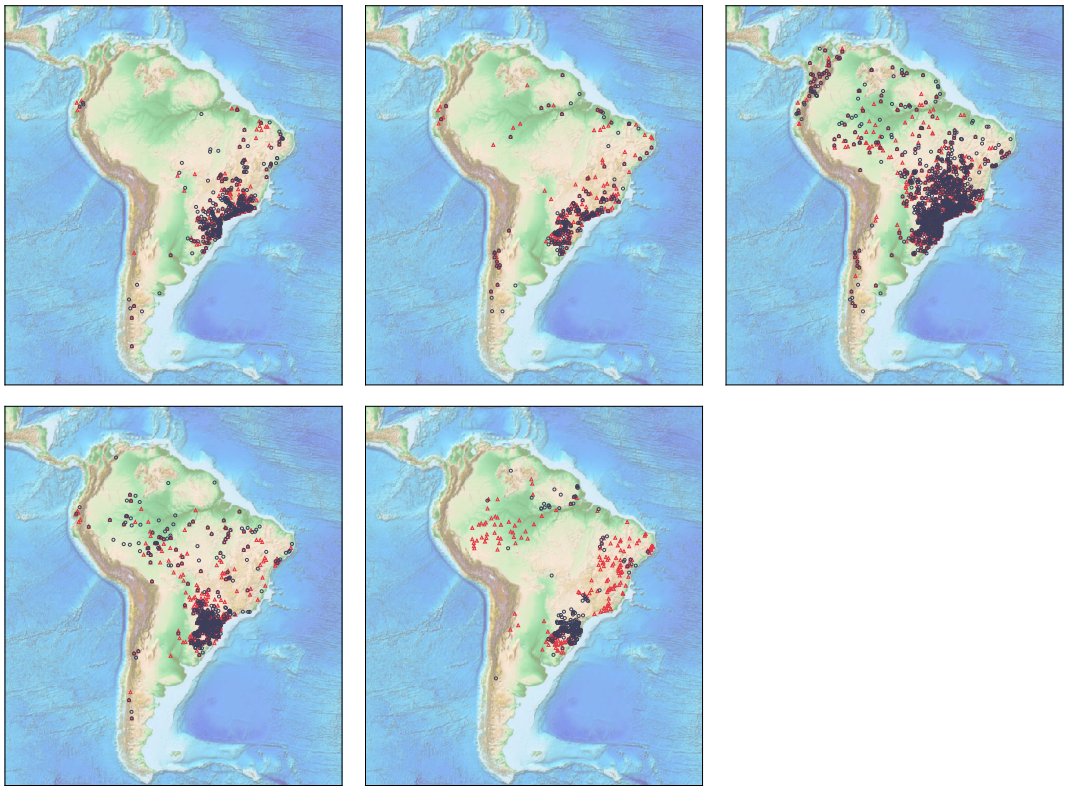
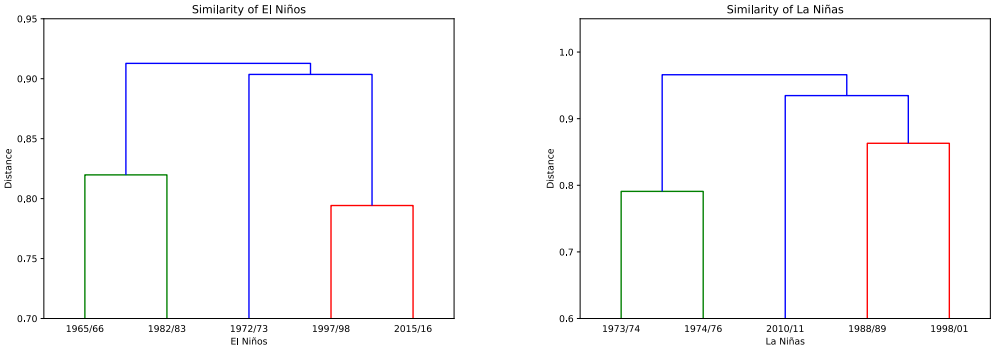


Figure 4: Impact pattern of individual El Niño events in a chronological order from the earliest on the top left panel to the most recent strong event on the bottom. Stations marked in red are the ones showing top ten peak river discharge during the active El Niño time of the respective event under the raw GSIM dataset, stations in blue with the dataset of inferred stations. Events depicted in the top row: 1965/66, 1972/73 and 1982/83, bottom row: 1997/98 and 2015/16 from left to right.

caused an economical damage that reached the billions in some regions of the continent, for example in the Itajaí-Açu basin or Argentina and moreover many thousand people were displaced as societies were hit unprepared by the extent of this flood (Fleischmann et al., 2020). In terms of precipitation, 1983 was also exceptionally strong and in some regions in the southeast of the continent it was the year with the most extreme rainfalls between 1980 and 2015 (Fleischmann et al., 2020). Its largest rainfall volumes were observed in the months of February, June and July, which has lead to a high water storage, soil saturation and even swamping in the affected areas. Those heavy antecedent rainfalls enhanced the magnitude of the floods. Regarding the water export to the South Atlantic in the region between São Francisco and the La Plata river outlets, 1983 reached an extraordinary high anomaly of 3.7, which is the largest value between 1980 and 2015, followed by 1998 with a value of 1.9 and 1992 with an anomaly of 1.1 (Fleischmann et al., 2020). According to the model predictions of Fleischmann et al. (2020) the continental water export to the oceans in 1983 was significantly lower and in fact below average, as the Amazon river, that contributes large parts to the continental water export,

faced a drought during this time. They also discuss anthropogenic factors that possibly contributed to the extent of the floods, such as the construction and activation of dams. Also the change of land use could possibly enhance such an extreme event, as the southeast of the continent shows a large scale transition from forest towards grassland and some regions record a loss of forest fraction and an increase in cropland fraction up to 30% (Yang et al., 2019). As cropland monocultures are known to have lower infiltration rates than natural forests, they provide lesser flood protection and make the region therefore more vulnerable to floods, which might have contributed to the magnitude of the 1982/83 event. The impact map of 1982/83 also has the farthest spread towards the north and towards the west of the continent of all El Niño events investigated. Its impact on the Colombian and Ecuadorian Andes was also significantly greater than for any other El Niño event in our study. An individual assessment of why 1982/83 was that intense and how individual components contributed to the flooding during the course of this El Niño requires further investigation.

3.4 Similarity of strong ENSO events



(a) Dendrogram of the five strongest El Niño events

(b) Dendrogram of the five strongest La Niña events

Figure 5: Results of the hierarchical clustering on the Jaccard-distances of the stations that show top ten peak river discharge during the active time of the respective ENSO event in the form of a dendrogram.

The results of the hierarchical clustering and the respective dendrograms are depicted in Figure 5. The tables S1 and S2 report the corresponding distance matrices for El Niño and La Niña events. The similarity analysis of the El Niño events resulted in three groups for a distance threshold of 0.82. The respective map with affected stations for each category is shown in Figure 6.
The first group with the highest similarity consists of the El Niño events of 1997/98 and 2015/16. According to Yu and Kim (2013) and Paek et al. (2017) the event of 1997/98 was a clear and strong Eastern Pacific (EP) El Niño and 2015/16 is considered a mixed El Niño that has features of both an Easter

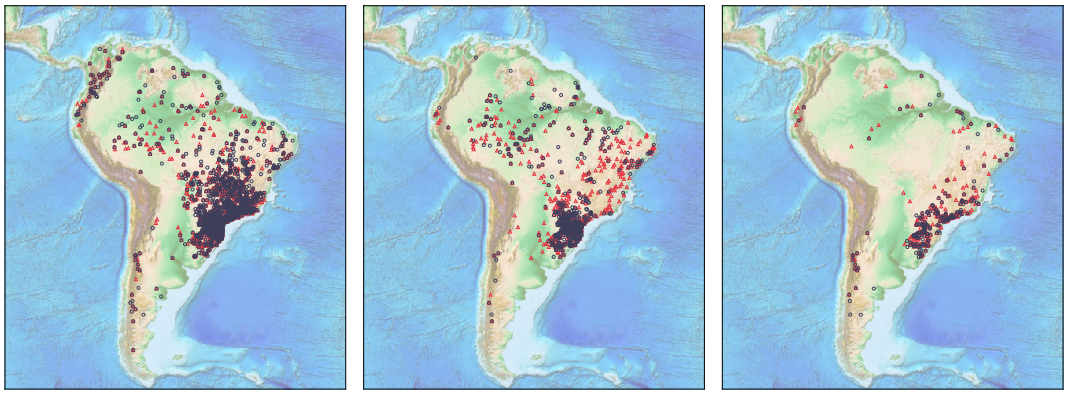


Figure 6: Impact pattern of the categories defined by the similarity analysis for El Niño. The maps are ordered as follows: left: 1965/66 & 1982/83, middle: 1997/98 & 2015/16, right: 1972/73. Stations marked in red are the ones showing top ten peak river discharge during the active El Niño time of the respective event under the raw GSIM dataset, stations in blue with the dataset of inferred stations.

and Central Pacific (CP) El Niño. Furthermore, they state that 1997 is the strongest EP El Niño and 2015 the strongest mixed El Niño to date, which makes them the strongest El Niño in their respective category. Both events reached a similar sea surface temperature (SST) anomaly of 3.5°C and had a similar evolution of the SST anomalies from the coast of South America to the international date line, however the dynamics of 2015 differ from 1997 in a way that it was dominated by the CP El Niño dynamics as well (Paek et al., 2017). Although their impact on U.S. climate was quite different, they impacted river discharge in South America in a similar manner.

The next group consists of the mixed El Niño of 1965/66 and the clear EP El Niño of 1982/83. The latter is the event that affected the most stations of all ENSO events investigated and induced heavy flooding events in the south of the United States, in Ecuador and the south east of Brazil (Fleischmann et al., 2020). The amount of affected stations during 1965/66 is with 353 stations only a little more than one third of the 1040 stations that showed top ten peak river discharge. During the El Niño period of 1982/83 however they both impacted similar regions and stations. 1965 shows a more concentrated impact on the southeastern Atlantic coast of Brazil and some scattered stations in the Patagonia region, Northern Ecuador and the Northern Atlantic coast of Brazil as well as the Tocantis-Araguaia basin. The event of 1982/83 on the other hand has a large and concentrated amount of stations along the south and southeast Atlantic coast of Brazil and the La Plata basin. Also, more stations in the Colombian and Ecuadorian Andes, in Patagonia and especially in the Amazon basin were affected.

The third category is the single event of 1972/73 which has dynamics of the EP type but is considered as a mixed El Niño according to the pattern correlation method by Paek et al. (2017). With 184 stations affected during the active phase of the 1972/73 El Niño it is the event with the fewest stations of all five

events investigated. The most stations are also concentrated along the south east of Brazil. However, stations seem to be more scattered and spread farther towards the north of Brazil, the Amazon basin and Patagonia.

The following paragraph describes the results of the similarity analysis for La Niña and its respective categories. For a cutoff at 0.87 we obtained three categories whose impact maps are presented in Figure 7. The events of 1973/74 and 1974/76 that are the most similar according to our similarity analysis occurred with very little temporal offset from each other. As two periods (JAS and ASO) of the three month running mean were below the threshold of -0.5 they are considered as two individual events. The temporal closeness is a potential explanation for the closeness of those events as already saturated soils and rivers with high discharge after the first event are likely to flood again more easily after only little amounts of precipitation, during the second event. The second group consists of the 1988/89 and 1998/01 events which are widely scattered north of 20°S latitude, and holds a more concentrated region in the south east of Brazil, which is unusual as this is not a region where La Niña normally induces a wetter climate. Another more concentrated region of stations is the northeast Brazilian Atlantic Coast where this region is among the most concentrated in terms of the overall distribution of stations as depicted in the Heatmap in Figure S1. Nevertheless this region as well as the whole north of South America typically has a wetter climate during a La Niña. In general, the impact pattern of the second category is similar to the first but intensified as a larger amount of stations were affected during the events of the second category. They both show a generally scattered pattern around the whole continent with two more concentrated regions. The event that shows the least similarity to all the other events is the La Niña of 2010/11. Although it led to intense flooding in some regions around the world like the Pakistan floods in 2010 or the 2010/11 Queensland floods, it is the event that affected the fewest stations in our analysis. Especially very few inferred stations were affected but are located in the same regions as in the first two categories. The comparison of the patterns between the inferred stations and the raw GSIM dataset of 2010/11 emphasizes the previously described temporal bias when using raw data, as the impact of this event seems to be much larger as to when using data that covers the whole timespan.
As the absolute distance values of the events are close together we performed a 100-fold cross validation, to backup the results of our similarity analysis. The outcome of the cross validation is depicted in Figure 8 as a graph, where the edge weights represent the percentage of those two events being grouped together during an iteration. The groups are consistent throughout every iteration and result in the same categories, which makes our results quite robust.

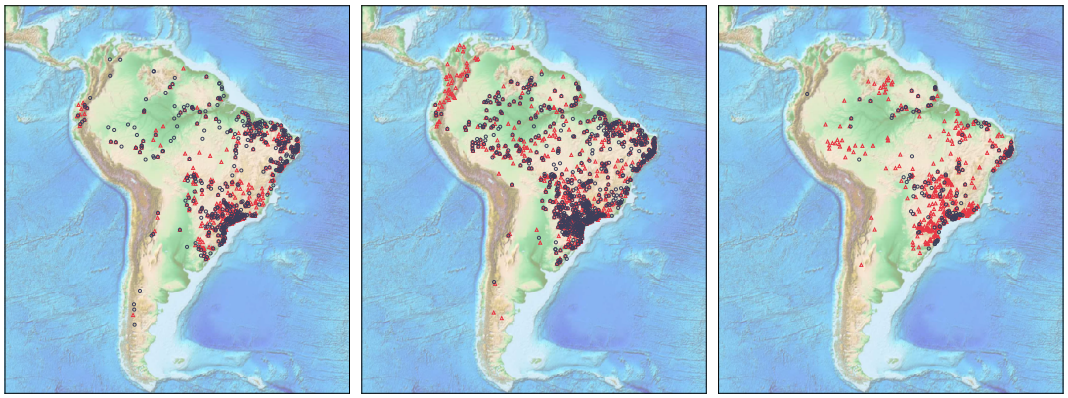


Figure 7: Impact pattern of the categories defined by the similarity analysis for La Niña. The maps are ordered as follows: left: 1973/74 & 1974/76, middle: 1988/89 & 1998/01, right: 2010/11. Stations marked in red are the ones showing top ten peak river discharge during the active El Niño time of the respective event under the raw GSIM dataset, stations in blue with the dataset of inferred stations.

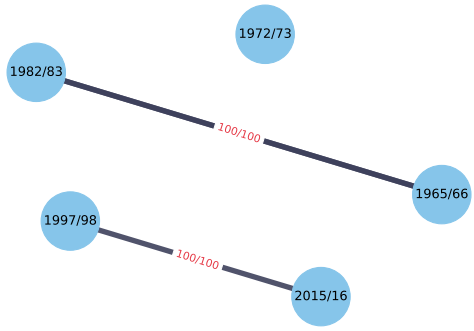


Figure 8: During each iteration i of the 100-fold cross-validation the cutoff level has been chosen such that we obtain 3 categories. For when two events e_i and e_j have been grouped together the entry e_i, e_j in the adjacency matrix has been incremented. The figure shows the respective graph for the adjacency matrix. The edge weights are labeled according to the score in the adjacency matrix.

3.5 Impact of different El Niño types

The El Niño types are known to have different impacts, such as on U.S. climate (Yu et al., 2012) or on East Asian summer precipitation (Wen et al., 2020). Regarding the impact pattern of river discharge in South America during the three mixed El Niños of 1965, 1972 and 2015 and the clear EP El Niños of 1982 and 1997, differences could be ascertained as well. The EP events are the ones with the largest number of affected stations. Together they impacted more stations than the three mixed events combined. Also larger parts in the Amazon basin were affected during EP events. The difference in the Amazon basin becomes even clearer when respecting the inferred stations only, as most stations in the Amazon basin during CP El Niños are stations that do not cover the whole timespan. Based on our small sample size it seems that EP El Niños have a more intense impact on river discharge than CP or mixed El Niños and that their impact extends over the whole continent, even to regions that are usually linked to a drier climate such as the Amazon basin. Also, the

Colombian and Ecuadorian Andes seem to be more affected by EP El Niños than by other El Niño types. The mixed El Niños on the other hand seem to be more concentrated on the southeast Atlantic Coast and along the northeastern part of the Atlantic Coast.

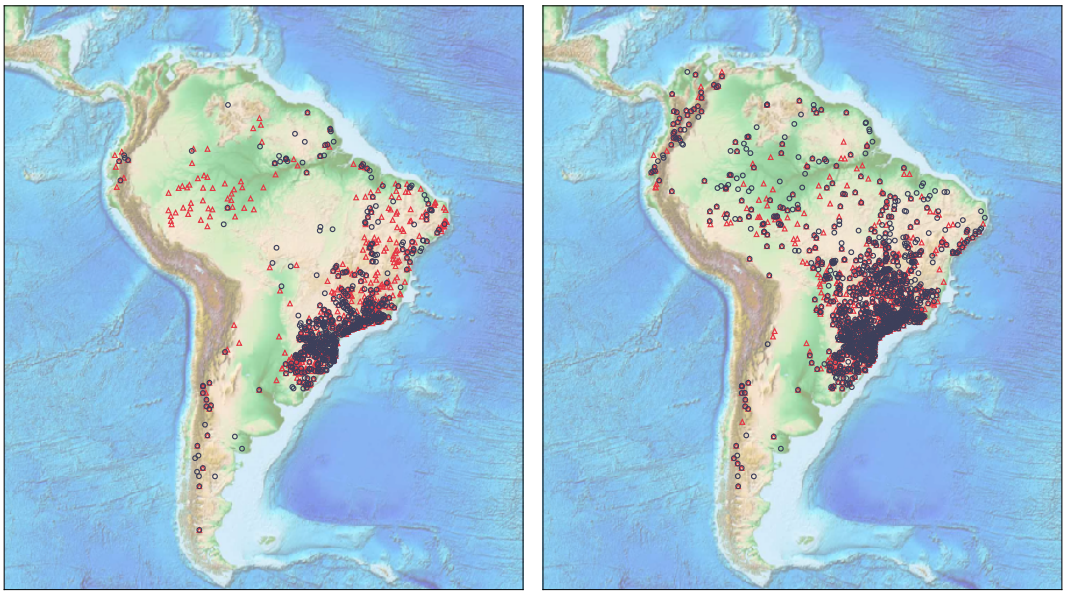


Figure 9: Impact pattern of stations that show top ten peak river discharge during Eastern Pacific El Niño (1982/83 and 1997/98) on the left map and central Pacific or mixed El Niño (1965/66, 1972/73 and 2015/16) on the right

The five strongest La Niña events investigatedbelong all to the same category following the dynamics of a CP La Niña event (Yuan and Yan, 2013). This is why a distinction between the La Niña types was not further investigated.

4 Conclusion

This study presents an approach under which large-scale hydrological analyses using in-situ data becomes feasible. We were able to overcome the spatio-temporal trade-off which in general constraints hydrological data-driven studies. By using Gaussian Process regression the amount of stations we could include in our study could be extended by 11-fold, which allowed us to go beyond the scale of basins or countries. The extension of the dataset allowed us to undertake a more in-depth analysis on river discharge over the whole continent of South America and its pattern of peak river discharge during strong ENSO events. However, limitations still remain, as conclusions can only be drawn for stations, which initially observed data and that provided the minimal overlap with our timespan. For Peru, Bolivia and the region of Patagonia, fewer stations are available in comparison to other countries and regions in South America which are therefore underrepresented in the GSIM dataset, and thus in our study. For those regions, a conclusion on the impact of strong ENSO events on peak river discharge could not be made. We hope that more institutes and organisations contribute to open and accessible data and share hydrological data with projects like the GSIM in the near future. And moreover we hope tha the installation and maintenance of gauging stations is being promoted such that the amount of missing data in the hydrological context is reduced naturally.
Another limitation of our study is the selection bias which is induced by the stations that initially cover the timespan. As shown in Figure 1 the stations are mostly located in the east of South America and spread along the Atlantic Coast from north to south. Only two stations are within the Amazon basin and none in the region of Colombia or Chile. As these timeseries are the base on which we train our models and infer data, the dynamics of those stations are partially induced into the timeseries for which we want to fill missing values, which might cause an inductive bias in our target time series. As a consequence of the limited amount of stations that initially cover the whole timespan, it is possible that some stations might not correlate strongly with any of the 216 stations. However it is possible that stations are teleconnected and highly correlated although they are not close to each other in space. On the other hand spatial proximity does not necessarily imply similar streamflow behaviour. It is not seldom the case that time series of gauging station in the same river show different behaviour and dynamics, as river discharge can drastically change during its travel time. So although in theory spatial proximity does not necessarily imply similar dynamics in the hydrological context, more evenly distributed stations as well as more stations in general would reduce the selection bias in our study.
The maps according to the El Niño types indicate that clear and strong EP El Niños have a stronger impact on the continent in terms of peak river discharge

and floods than events that belong to the type of mixed El Niños. EP El Niños are capable of impacting the Amazon basin as well as large parts in the north of the continent that are usually associated with drier climate during an active El Niño. The underlying mechanisms of why EP events tend to have a larger impact than the other categories is beyond the scope of this work and needs further investigation. Also the underlying causes and driving forces of the strong impact on river discharge during the 1982/83 El Niño needs to be further investigated, as this event outnumbered every other El Niño or La Niña in terms of affected stations. Also its spatial impact area extended over the whole continent both in raw and inferred stations, like no other event. Disentangling the underlying causality that has led to this strong event and why it caused such a strong impact on river discharge and floods could enhance flood risk forecasting to better assess extreme events driven by the ENSO.

Stations in areas that are prone to higher ENSO-driven precipitation anomalies are also those regions where peak river discharge was ascertained and often shows the highest concentration in stations. Yet, our analysis shows that high streamflow extends to large regions above this precipitation-prone region and farther north (south) for El Niños (La Niñas). El Niño enhances and impacts river discharge in the southeast of Brazil and Uruguay, but also rivers in the Amazon basin, the Tocantins-Araguaia basin, the São Francisco basin or the south of Chile can possibly be affected by strong events. El Niño and La Niña events are generally capable of impacting the continent in the same regions. For La Niña events, a high concentration of affected stations is along the northeastern Atlantic Coast of Brazil , which is more prominent during La Niña rather than El Niño, as well as in the southeast of Brazil. The latter is also the region where El Niño impacted the continent the most. Also large parts of the Amazon basin were subject to peak precipitation and floods during the ENSO which is in logical relation for La Niña but unexpected for El Niños. The large spread of affected stations corroborates the nonlinearity between precipitation and floods identified by Stephens et al. (2015).

This study can easily be carried forward and applied to other regions such as North America, Asia or Africa, or even an analysis on a global scale to further investigate ENSO impact on river discharge. The latter would be especially interesting as ENSO also has a worldwide impact. Also under the aspect of revealing global teleconnections between river segments, a global analysis would be especially interesting.

From a technical perspective, this approach could be further enhanced by a more dynamical approach to estimate missing values, such that we do not rely on stations that cover the full timespan. Multiple models could be trained for one target time series that only cover parts of the timespan. This would allow us to consider more stations in our basis on which we can estimate missing values. Also, model predictions for smaller time spans would better capture a change of river dynamics as the highest correlated time series might change

from time segment to time segment, just like the characteristics of rivers can change over time due to natural or anthropogenic causes.

To summarize, there are multiple findings we can derive from this study. First, inferring hydrological streamflow data with Gaussian Processes is a feasible approach and yields good results. Second, ENSO events are capable of impacting rivers on a spatially larger scale then ENSO's impact on precipitation and temperature anomalies would suggest. Third, from the pattern of affected stations we noticed that Eastern Pacific El Niños have a stronger drive towards the north of the continent and the Amazon basin than mixed El Niños and have a potentially stronger impact. The 1982/83 Eastern Pacific El Niño was extraordinary strong in comparison to other events as it impacted three times more stations than any other El Niño, although it was not the event which achieved the highest value on the index scale. It is meaningful to group and categorize ENSO events and estimate the magnitude of their impact based on their index, but there is no such thing as two identical ENSO events. Even if they have the same physical background and comparable value on the index scale, they might still show very different behaviour. As the 1982/83 event showed, the index values is a good indicator but no guarantor to assess its magnitude correctly, as events with smaller values might have larger impacts in the end. Disentangling the interplay of individual features, the driving forces of ENSO events and their respective impact on river discharge therefore remains a research field of great interest for enhancing hydrological forecasting and to develop a better understanding of why each event is so unique and different from the other.

References

Lorenzo Alfieri, Berny Bisselink, Francesco Dottori, Gustavo Naumann, Ad de Roo, Peter Salamon, Klaus Wyser, and Luc Feyen. Global projections of river flood risk in a warmer world: RIVER FLOOD RISK IN A WARMER WORLD. *Earth's Future*, 5(2):171–182, February 2017. ISSN 23284277. doi: 10.1002/2016EF000485. URL http://doi.wiley.com/10.1002/2016EF000485.

J. Dietrich, A. H. Schumann, M. Redetzky, J. Walther, M. Denhard, Y. Wang, B. Pfützner, and U. Büttner. Assessing uncertainties in flood forecasts for decision making: prototype of an operational flood management system integrating ensemble predictions. *Natural Hazards and Earth System Sciences*, 9(4):1529–1540, August 2009. ISSN 1684-9981. doi: 10.5194/nhess-9-1529-2009. URL https://nhess.copernicus.org/articles/9/1529/2009/.

Maxx Dilley and Barry N. Heyman. ENSO and Disaster: Droughts, Floods and El Niño/Southern Oscillation Warm Events. *Disasters*, 19(3):181–193, September 1995. ISSN 03613666. doi: 10.1111/j.1467-7717.1995.tb00338.x. URL http://doi.wiley.com/10.1111/j.1467-7717.1995.tb00338.x.

Hong Xuan Do, Lukas Gudmundsson, Michael Leonard, and Seth Westra. The Global Streamflow Indices and Metadata Archive (GSIM) – Part 1: The production of a daily streamflow archive and metadata. *Earth System Science Data*, 10(2):765–785, April 2018. ISSN 1866-3516. doi: 10.5194/essd-10-765-2018. URL https://essd.copernicus.org/articles/10/765/2018/.

Esmaeel Dodangeh, Vijay P. Singh, Binh Thai Pham, Jiabo Yin, Guang Yang, and Amirhosein Mosavi. Flood Frequency Analysis of Interconnected Rivers by Copulas. *Water Resources Management*, 34(11):3533–3549, September 2020. ISSN 0920-4741, 1573-1650. doi: 10.1007/s11269-020-02634-0. URL http://link.springer.com/10.1007/s11269-020-02634-0.

R. Emerton, H. L. Cloke, E. M. Stephens, E. Zsoter, S. J. Woolnough, and F. Pappenberger. Complex picture for likelihood of ENSO-driven flood hazard. *Nature Communications*, 8(1):14796, April 2017. ISSN 2041-1723. doi: 10.1038/ncomms14796. URL http://www.nature.com/articles/ncomms14796.

Ayan Santos Fleischmann, Vinícius Alencar Siqueira, Sly Wongchuig-Correa, Walter Collischonn, and Rodrigo Cauduro Dias De Paiva. The great 1983 floods in South American large rivers: a continental hydrological modelling approach. *Hydrological Sciences Journal*, 65(8):1358–1373, June 2020. ISSN

0262-6667, 2150-3435. doi: 10.1080/02626667.2020.1747622. URL https://www.tandfonline.com/doi/full/10.1080/02626667.2020.1747622.

GPy. GPy: A gaussian process framework in python. http://github.com/SheffieldML/GPy, 2012.

Gabriela Guimarães Nobre, Sanne Muis, Ted I. E. Veldkamp, and Philip J. Ward. Achieving the reduction of disaster risk by better predicting impacts of El Niño and La Niña. *Progress in Disaster Science*, 2:100022, July 2019. ISSN 2590-0617. doi: 10.1016/j.pdisas.2019.100022. URL https://www.sciencedirect.com/science/article/pii/S2590061719300225.

Yukiko Hirabayashi, Roobavannan Mahendran, Sujan Koirala, Lisako Konoshima, Dai Yamazaki, Satoshi Watanabe, Hyungjun Kim, and Shinjiro Kanae. Global flood risk under climate change. *Nature Climate Change*, 3 (9):816–821, September 2013. ISSN 1758-678X, 1758-6798. doi: 10.1038/nclimate1911. URL http://www.nature.com/articles/nclimate1911.

S. N. Jonkman. Global Perspectives on Loss of Human Life Caused by Floods. *Natural Hazards*, 34(2):151–175, February 2005. ISSN 0921-030X, 1573-0840. doi: 10.1007/s11069-004-8891-3. URL http://link.springer.com/10.1007/s11069-004-8891-3.

Nathan J. L. Lenssen, Lisa Goddard, and Simon Mason. Seasonal Forecast Skill of ENSO Teleconnection Maps. *Weather and Forecasting*, 35(6):2387–2406, December 2020. ISSN 0882-8156, 1520-0434. doi: 10.1175/WAF-D-19-0235.1. URL https://journals.ametsoc.org/view/journals/wefo/35/6/WAF-D-19-0235.1.xml.

M. J. McPhaden, S. E. Zebiak, and M. H. Glantz. ENSO as an Integrating Concept in Earth Science. *Science*, 314(5806):1740–1745, December 2006. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1132588. URL https://www.sciencemag.org/lookup/doi/10.1126/science.1132588.

Michael J. McPhaden, editor. *El Niño southern oscillation in a changing climate*. Geophysical monograph series. Wiley-American Geophysical Union, Hoboken, NJ, first edition edition, 2020. ISBN 9781119548126.

Daisuke Nohara, Akio Kitoh, Masahiro Hosaka, and Taikan Oki. Impact of Climate Change on River Discharge Projected by Multimodel Ensemble. *Journal of Hydrometeorology*, 7(5):1076–1089, October 2006. ISSN 1525-7541, 1525-755X. doi: 10.1175/JHM531.1. URL http://journals.ametsoc.org/doi/10.1175/JHM531.1.

NOOA's Climate Prediction Center. 2021. URL https://origin.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_v5.php.

Houk Paek, Jin-Yi Yu, and Chengcheng Qian. Why were the 2015/2016 and 1997/1998 extreme El Niños different?: Contrasting 1997/1998 and 2015/2016 El Niños. *Geophysical Research Letters*, 2017. ISSN 00948276. doi: 10.1002/2016GL071515. URL http://doi.wiley.com/10.1002/2016GL071515.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass, 2006. ISBN 9780262182539. OCLC: ocm61285753.

Hannah Ritchie and Max Roser. Natural Disasters. *Our World in Data*, June 2014. URL https://ourworldindata.org/natural-disasters.

Jochen Schanze, Evzen Zeman, and J. Marsalek, editors. *Flood risk management: hazards, vulnerability and mitigation measures*. Number vol. 67 in NATO science series. Series IV, Earth and environmental sciences. Springer, Dordrecht, 2006. ISBN 9781402045967 9781402045974 9781402045981. OCLC: ocm71425441.

J. Q. Shi, B. Wang, R. Murray-Smith, and D. M. Titterington. Gaussian Process Functional Regression Modeling for Batch Data. *Biometrics*, 63 (3):714–723, September 2007. ISSN 0006341X. doi: 10.1111/j.1541-0420.2007.00758.x. URL http://doi.wiley.com/10.1111/j.1541-0420.2007.00758.x.

E. Stephens, J. J. Day, F. Pappenberger, and H. Cloke. Precipitation and floodiness. *Geophysical Research Letters*, 42(23), December 2015. ISSN 0094-8276, 1944-8007. doi: 10.1002/2015GL066779. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/2015GL066779.

Alexander Y. Sun, Dingbao Wang, and Xianli Xu. Monthly streamflow forecasting using Gaussian Process Regression. *Journal of Hydrology*, 511:72–81, April 2014. ISSN 00221694. doi: 10.1016/j.jhydrol.2014.01.023. URL https://linkinghub.elsevier.com/retrieve/pii/S0022169414000298.

USGS.gov. Surface and overland water runoff. *Surface and Overland Water Runoff*, June 2021. URL https://www.usgs.gov/special-topic/water-science-school/science/runoff-surface-and-overland-water-runoff?qt-science_center_objects=0#qt-science_center_objects. publisher: usgs.gov.

Michelle T.H. van Vliet, Wietse H.P. Franssen, John R. Yearsley, Fulco Ludwig, Ingjerd Haddeland, Dennis P. Lettenmaier, and Pavel Kabat. Global river discharge and water temperature under climate change. *Global Environmental Change*, 23(2):450–464, April 2013. ISSN 09593780. doi:

10.1016/j.gloenvcha.2012.11.002. URL https://linkinghub.elsevier.com/retrieve/pii/S0959378012001331.

Na Wen, Laurent Li, and Jing-Jia Luo. Direct impacts of different types of El Niño in developing summer on East Asian precipitation. *Climate Dynamics*, 55(5-6):1087–1104, September 2020. ISSN 0930-7575, 1432-0894. doi: 10.1007/s00382-020-05315-1. URL https://link.springer.com/10.1007/s00382-020-05315-1.

Dai Yamazaki, Satoshi Watanabe, and Yukiko Hirabayashi. Global Flood Risk Modeling and Projections of Climate Change Impacts. In Guy J-P. Schumann, Paul D. Bates, Heiko Apel, and Giuseppe T. Aronica, editors, *Geophysical Monograph Series*, pages 185–203. John Wiley & Sons, Inc., Hoboken, NJ, USA, June 2018. ISBN 9781119217886. doi: 10.1002/9781119217886.ch11. URL http://doi.wiley.com/10.1002/9781119217886.ch11.

Yan, Huan Wu, Guojun Gu, Philip J. Ward, Lifeng Luo, Xiaomeng Li, Zhijun Huang, and Jing Tao. Exploring the ENSO Impact on Basin-Scale Floods Using Hydrological Simulations and TRMM Precipitation. *Geophysical Research Letters*, 47(22), November 2020. ISSN 0094-8276, 1944-8007. doi: 10.1029/2020GL089476. URL https://onlinelibrary.wiley.com/doi/10.1029/2020GL089476.

Hui Yang, Chris Huntingford, Andy Wiltshire, Stephen Sitch, and Lina Mercado. Compensatory climate effects link trends in global runoff to rising atmospheric CO $_2$ concentration. *Environmental Research Letters*, 14(12):124075, December 2019. ISSN 1748-9326. doi: 10.1088/1748-9326/ab5c6f. URL https://iopscience.iop.org/article/10.1088/1748-9326/ab5c6f.

Jin-Yi Yu and Seon Tae Kim. Identifying the types of major El Niño events since 1870: TYPES OF MAJOR EL NIÑO EVENTS SINCE 1870. *International Journal of Climatology*, 33(8):2105–2112, June 2013. ISSN 08998418. doi: 10.1002/joc.3575. URL http://doi.wiley.com/10.1002/joc.3575.

Jin-Yi Yu, Yuhao Zou, Seon Tae Kim, and Tong Lee. The changing impact of El Niño on US winter temperatures: IMPACT OF EL NINO ON US TEMPERATURES. *Geophysical Research Letters*, 39(15), August 2012. ISSN 00948276. doi: 10.1029/2012GL052483. URL https://onlinelibrary.wiley.com/doi/10.1029/2012GL052483.

Yuan Yuan and HongMing Yan. Different types of La Niña events and different responses of the tropical atmosphere. *Chinese Science Bulletin*, 58(3):406–415, January 2013. ISSN 1001-6538, 1861-9541. doi:

10.1007/s11434-012-5423-5. URL http://link.springer.com/10.1007/s11434-012-5423-5.

Matija Zorn. Natural Disasters and Less Developed Countries. In Stanko Pelc and Miha Koderman, editors, *Nature, Tourism and Ethnicity as Drivers of (De)Marginalization*, volume 3, pages 59–78. Springer International Publishing, Cham, 2018. ISBN 9783319590011 9783319590028. doi: 10.1007/978-3-319-59002-8_4. URL http://link.springer.com/10.1007/978-3-319-59002-8_4.

5 Supplementary

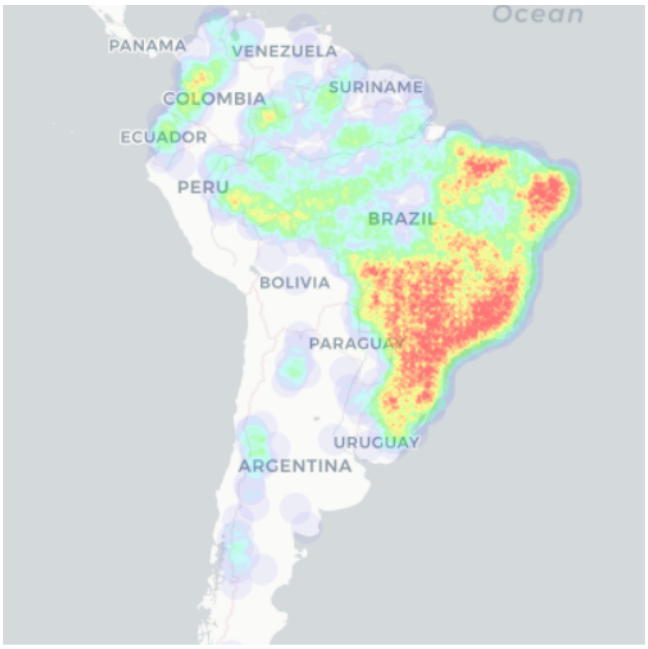


Figure S1: Heatmap indicating the density of available stations in the GSIM dataset, where the area of south Brazil up until 20° latitudinal and the north east of the Brazilian Atlantic coast are the regions of highest density. For the Amazon basin and the Colombian and Ecuadorian Andes density of stations is rather low, as well as in some patches of Argentina and Paraguay.

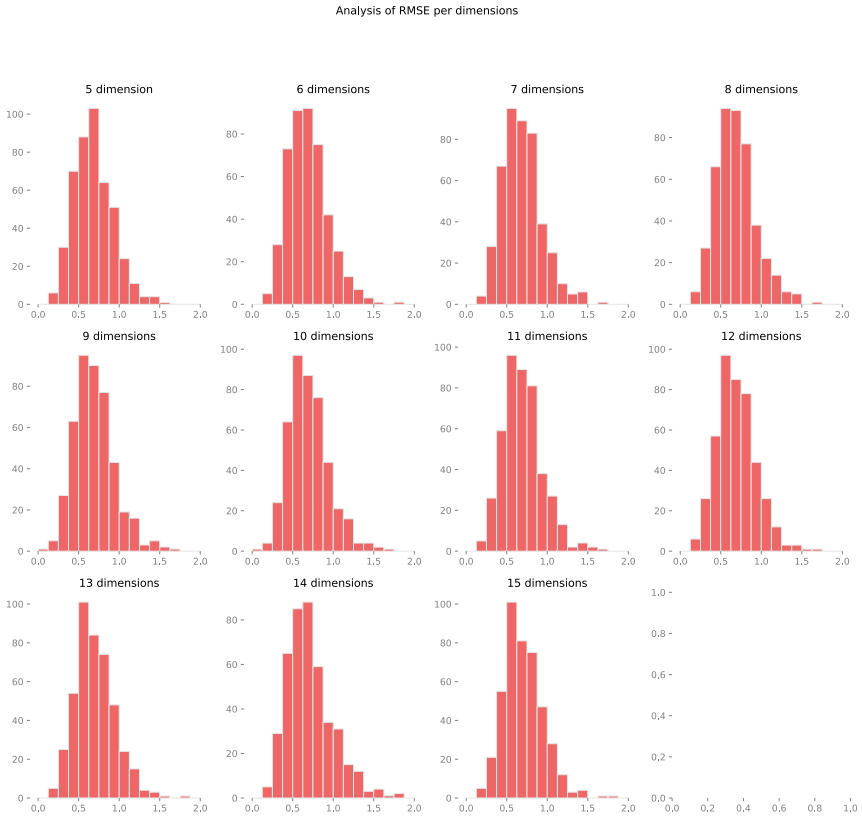


Figure S2: For a subsample of 500 timeseries we trained GP models to predict missing values on its 5 - 15 highest correlated timeseries and plotted for each number of timeseries included a histogram of RMSE values scored during testing. As the results for each iteration step were quite robust with little to no outliers the dimensionality of ten has been choosen as a target value of timeseries to include during training.

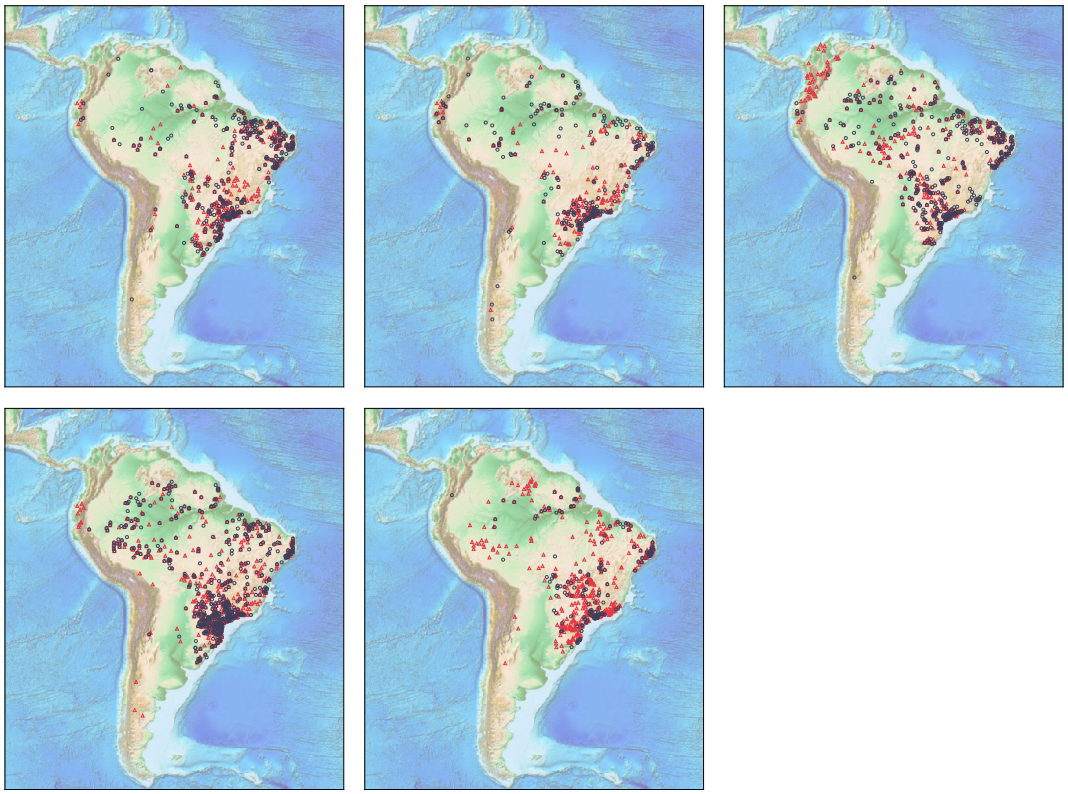


Figure S3: Impact pattern of individual La Niña events in a chronological order from the earliest on the top left panel to the most recent strong event on the bottom. Stations marked in red are the ones showing top ten peak river discharge during the active La Niña time of the respective event under the raw GSIM dataset, stations in blue with the dataset of inferred stations. Events depicted in the top row: 1973/74, 1974/76 and 1988/89, bottom row: 1998/01 and 2010/11 from left to right

	1965/66	1972/73	1982/83	1997/98	2015/16
1965/66	0.00	0.90	0.82	0.91	0.89
1972/73	0.90	0.00	0.91	0.88	0.90
1982/83	0.82	0.91	0.00	0.81	0.87
1997/98	0.91	0.88	0.81	0.00	0.79
2015/16	0.89	0.90	0.87	0.79	0.00

Table S1: Distance Matrix D of the strongest El Niño investigated

	1973/74	1974/76	1988/89	1998/01	2010/11
1973/74	0.00	0.79	0.87	0.94	0.97
1974/76	0.79	0.00	0.89	0.94	0.96
1988/89	0.87	0.89	0.00	0.86	0.93
1998/01	0.94	0.94	0.86	0.00	0.91
2010/11	0.97	0.96	0.93	0.91	0.00

Table S2: Distance Matrix D of the strongest La Niña investigated

Statement of Authorship

I hereby certify that I have written this Bachelor's thesis independently and only with the means indicated, and that all passages taken from other works in terms of wording or meaning have been marked by indicating the sources. This Bachelor's thesis has not been submitted in the same or a similar form as an examination in any other degree program.

Tübingen, 8th September 2021

Place, Date Signature