Predicting Droughts in the Amazon Basin based on Global Sea Surface Temperatures

Dario Lepke

September 09, 2022

```
## Lade nötiges Paket: sp
##
## Attache Paket: 'patchwork'
## Das folgende Objekt ist maskiert 'package:raster':
##
##     area
```

Introduction

Motivation

- The Amazon basin is a key hotspot of biodiversity, carbon storage and moisture recycling
- Hydrological extremes affect ecosystem and populations tremendously
- Droughts in the Amazon rainforest can have severe biomass carbon impact
- Severe Amazon drought in 2010 had total biomass carbon impact of 2.2 PgC, affected area $3miokm^2$

Related work

- Ciemer et al. (2020) established an early warning indicator for water deficits in the central Amazon basin (CAB)

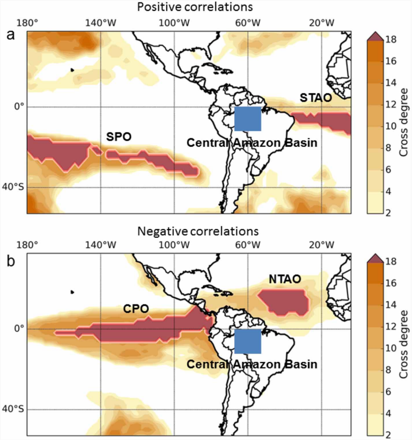


Figure 1: Cross degree between sea surface temperature and continental rainfall anomalies. For each grid cell of sea surface temperature in the

Our approach

- Inspect spatial and temporal characteristics in raw data
- Directly predict rain from SST
- Use lasso and fused lasso
- model evaluation with cross validation for time series

Explorative analysis

The data

- Rain data from CHIRPS ()
- CHIRPS contains in-situ and satellite data
- SST data from ERSST (Extended Reconstructed Sea Surface Temperature)
- ERSST is reanalysis of observation data (made by ships and buoys for example), missing data filled by interpolation techniques
- These are the same data sets as in Ciemer et al. (2020)

Explorative analysis Rain

- show area
- show mean and sd
- show glyph plots

Explorative analysis SST

- show mean and sd

Correlation analysis

Correlation analysis

- show timelag 0, raw and de-seasonalized

Clustering

Motivation/ Overview

- explorative analysis has shown spatial and temporal differences in the precipitation data
- we explored this further using k-means clustering
- steps: find optimal k via pca and gap statistic
- apply k-means to original precipitation data
- we compared k-means and k-medoid with and without PCA via the gap statistic
- here show only k-means with PCA as it gave best results
- applying the regression models to separate clusters might improve predictions
- Using 3 principal components and 5 cluster centers with k-means gave best results on gap statistic

k-means

- Our objective is to find k internally homogeneous and externally heterogeneous clusters
- Similarity is measured by the euclidean distance

$$d(x_i, x_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = ||x_i - x_{i'}||^2 (\#eq: eucl - dist) \quad (1)$$

- And we want to minimize the sum of distances inside all clusters, given by:

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} ||x_i - x_{i'}||^2 = \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - \bar{x}_k||^2 (\#eq: w \quad (2)$$

where $\bar{x} = (\bar{x}_{1k}, ..., \bar{x}_{pk})$ stands for the mean vectors of the kth

gap statistic

- number of clusters has to be defined beforehand
- we decided on the optimal number of k using the gap statistic
- Let W_k be $W(C)$ for fix k
- We compare W_k from the precipitation data with average W_k^* from B Monte Carlo sampled data sets

$$Gap(k) = E\{log(W*_k)\} - log(W_k).(\#eq:gap) \tag{3}$$

- We choose k as smallest k such that

$$Gap(k) \geq Gap(k+1) - s_{k+1} \tag{4}$$

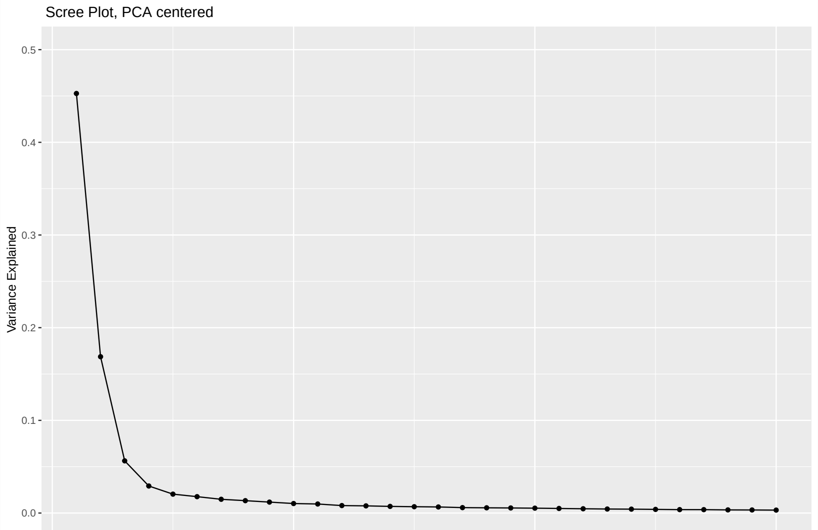
- s_{k+1} is $sd_k\sqrt{1 + 1/B}$, and sd the standard deviation of log(W*_k)

PCA

- Before running k-means we center the precipitation data and apply a PCA to reduce the large number of correlated variables to a few
- The new variables are linear combinations of the original variables
- Here: Each variable is a month of precipitation data in the CAB

Scree Plot, PCA after centering

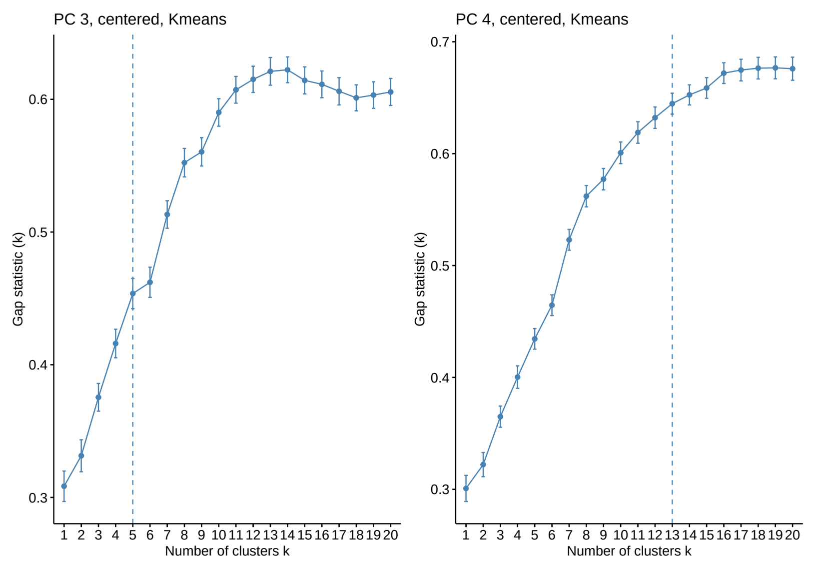
```
## Scale for 'y' is already present. Adding another scale f
## replace the existing scale.
```



Screeplot

- The “elbow” be observe in the screeplot suggest 3 or 4 principal components
- The first 3 and 4 first PC explain 67.77 and 70.79 of the variance respectively.
- We compare the gap statistic results for 3 and 4 PC

Gap statistic results



- The k-means gap statistic on the first 3 PC proposes 5

Clustering results

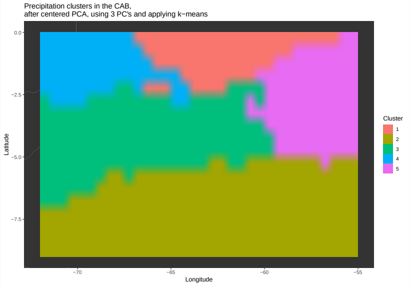


Figure 3: Spatial distribution of the found clusters in the CAB. We applied a centered PCA on the data and used 3 principal components before applying the k-means algorithm

- We find 5 clusters of different sizes
- The found clusters are almost completely spatially coherent although we did not include any spatial dependencies in the clustering

The lasso

Definition of the lasso

- We now consider the lasso regression problem

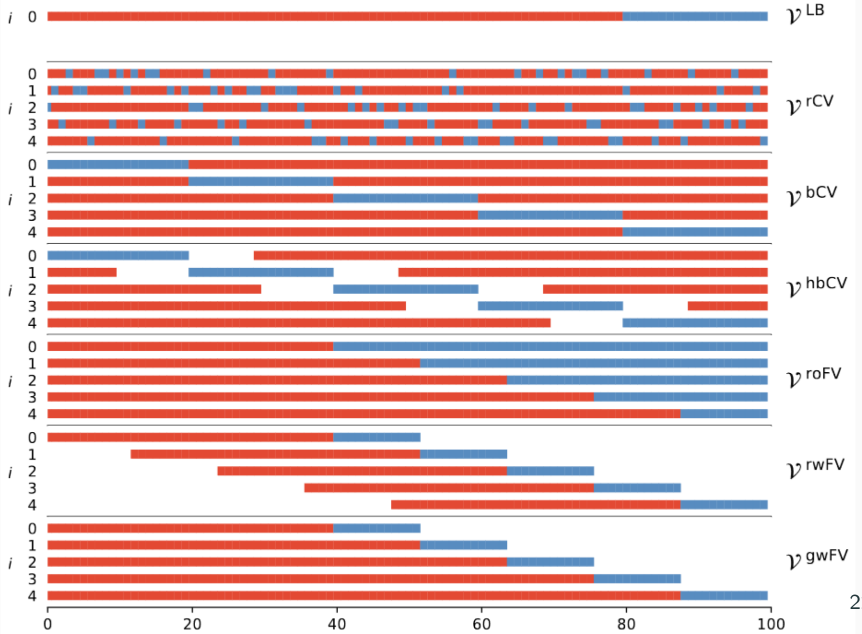
$$\min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1](\#eq:glmnet) \tag{5}$$

- In our setting n « p, so lasso is natural choice
- The problem is solved using coordinate descent
- Due to the time dependencies in our data normal Cross Validation may be unjustified

Model evaluation

- Our goal is to train a model that can also predict well on new, unseen data
- We simulate the situation of unseen data by splitting our data into one part for model selection and another part for model evaluation
- Model evaluation is usually done via Cross Validation, but classic Cross Validation does not take into account the time dependency in our data

Forward selection



Forward selection

- We compute a λ-vector for the complete training set
- For each fold we fit a model with this λ-vector
- We compute the prediction error for the cv-test set of each fold
- Choose λ_{min}, λ that minimizes average MSE over all folds
- Fit model on complete selection data with λ_{min} and compute MSE on evaluation data

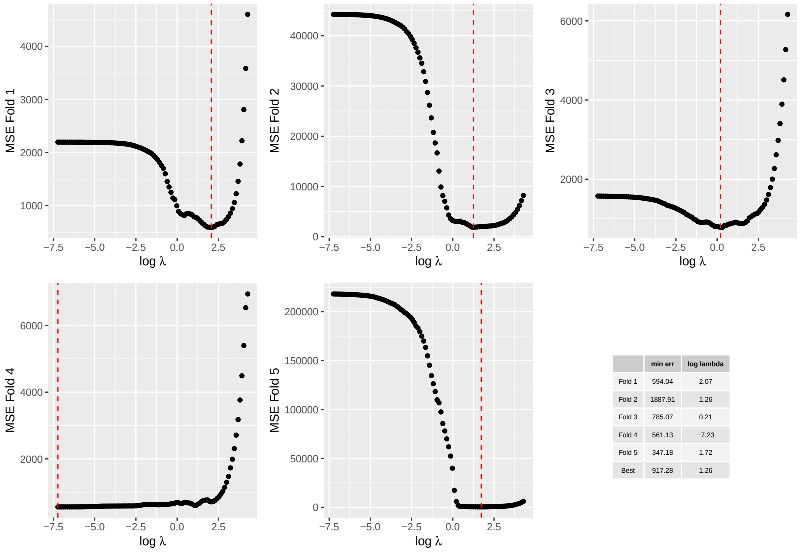
lasso settings

- lasso
- lasso with standardized features
- lasso with de-seasonalized SST
- lasso with differentiated SST
- lasso on clusters

lasso results TODO

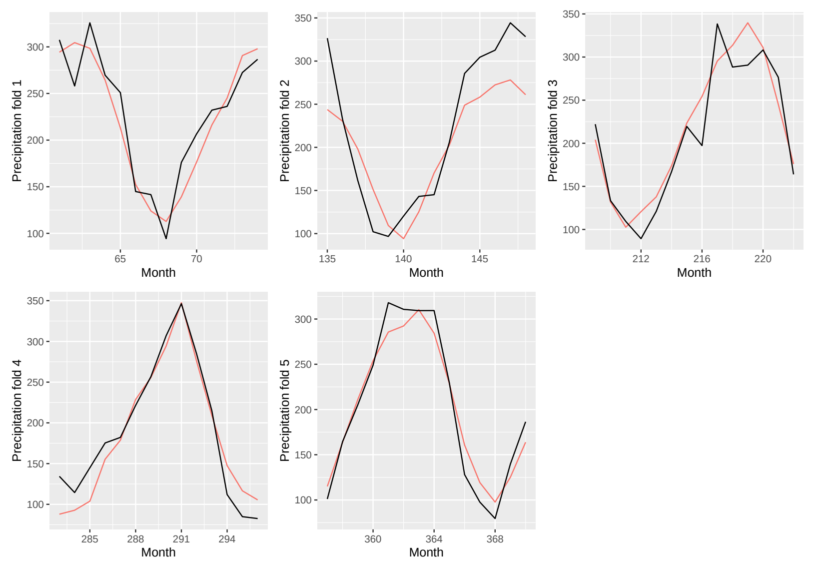
- Show only best model results
- lasso with standardized features
- show MSE in plots
- show predictions in plots
- show predictions
- show coefficients
- display table

MSE in each fold



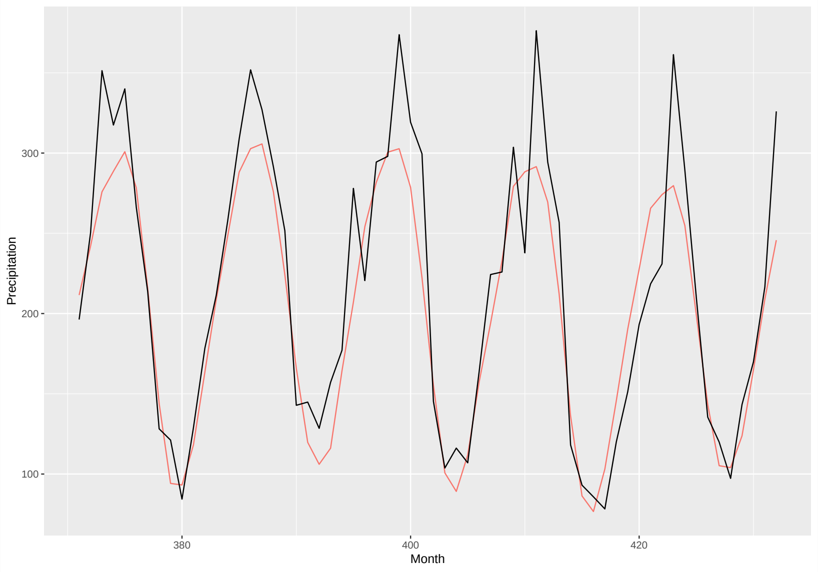
	min err	log lambda
Fold 1	594.04	2.07
Fold 2	1887.91	1.26
Fold 3	785.07	0.21
Fold 4	561.13	−7.23
Fold 5	347.18	1.72
Best	917.28	1.26

Predictions on test set, for each Fold



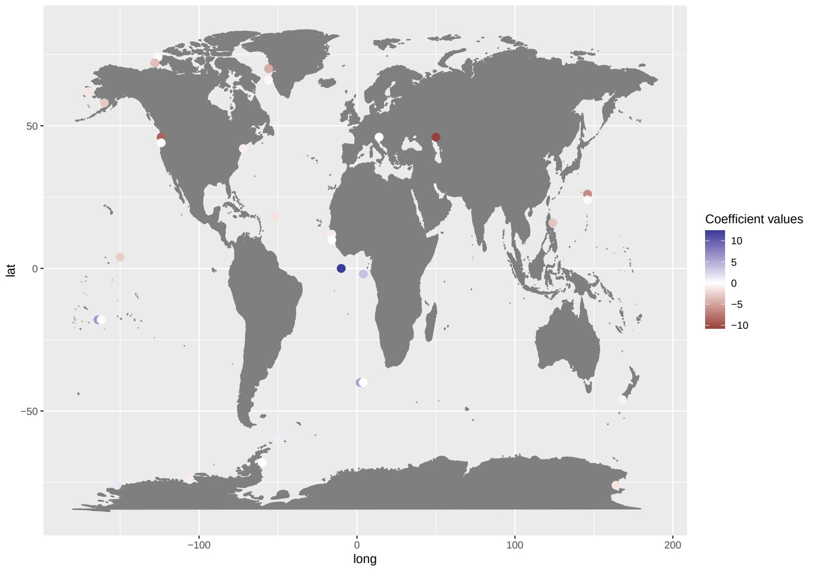
Predictions on External Test Set

Predictions on External Test Set




```
## [1] 1214.49
```

SST Regions chosen by the lasso



Lasso results all models

Lasso results all models

```
##          mse min lambda min
## stand    1214.49       3.52
## original 1314.93       3.52
## diff     1361.82       2.21
## deseas   1809.45       1.75
```

Lasso on clusters results

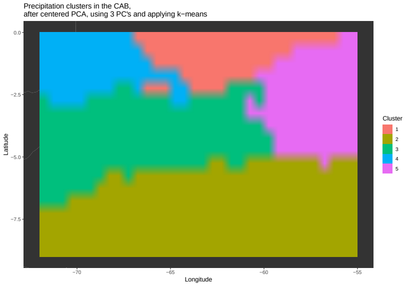
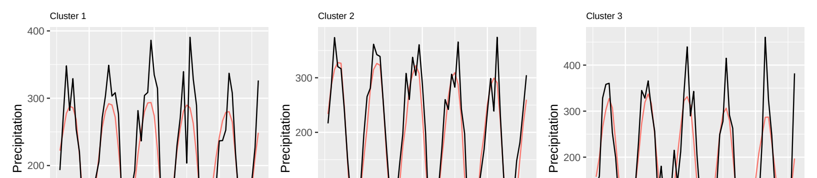


Figure 6: Spatial distribution of the found clusters in the CAB. We applied a centered PCA on the data and used 3 principal components before applying the k-means algorithm



Summary lasso results

- We compared different settings for the lasso
- Lasso, lasso with standardization, de-seasonalizing, differentiation and on clusters
- Lasso with standardized SST worked best
- Can predict general seasonality, but still fails to predict peaks in precipitation
- Clustering the CAB improves only on one cluster, but on this cluster peaks can be predicted better than in the unclustered model
- Lasso chooses single “points” and not whole areas
- The points chosen as coefficients differ in the models, and can be very far away from the CAB

Discussion Validation approach (maybe discuss this at the end)

- For the CAB we can not predict large values in the hold-out set, on cluster 2 it works a little better
- Possible explanations:
- Our validation approach works better when train and test set are similar in terms of seasonality and trend
- When train and test differ, predictions might not work so well (test of stationarity in folds)
- Differentiating and de-seasonalizing could not solve this problem
- Predictions work better when the precipitation remains fairly stable over time, see Cluster 2
- Final model uses complete model selection data, possibly some of that information is not useful anymore if it's toocfar away from hold-out time frame
- Our validation approach is a trade-off between efficient use of

The fused lasso

Definition of the fused lasso

- Fused lasso, “fuses” predictors together
- It penalizes the difference of close predictores
- Therefore close predictors should be similar

$$\min_{\beta} 1/2 \sum_{i=1}^{n} (y_i - x_i^T \beta_i)^2 + \lambda \sum_{i,j \in E} |\beta_i - \beta_j| + \gamma \cdot \lambda \sum_{i=1}^{p} |\beta_i|, (\#eq : fused-lasso) \tag{6}$$

- with x_i being the ith row of the predictor matrix and E is the edge set of an underlying graph.
- The third term $\gamma \cdot \lambda \sum_{i=1}^{p} |\beta_i|$, controls the sparsity of the coefficients.
- $\gamma = 0$ leads to complete fusion of the coefficients (no sparsity) and $\gamma > 0$ introduces sparsity to the solution, with higher values placing more priority on sparsity.

Fused lasso optimization

- Lets consider the problem in the notation of the generalizes lasso problem

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|D\beta\|_1, (\#eq : gen - lasso) \tag{7}$$

- where $y \in \mathbb{R}^n$ is the vector of the outcome, $X \in \mathbb{R}^{n \times p}$ a predictor matrix, $D \in \mathbb{R}^{m \times p}$ denotes a penalty matrix, and $\lambda \geq 0$ is a regularization parameter.
- The dual path algorithm solves not the primal but the dual solution of the problem and computes the solution for a whole path instead of single values of λ.

- Let's consider the case when $X = I$ and $rank(X) = p$ (this is called the "signal approximator" case), the dual problem of @ref(eq:gen-lasso) is then:

$$\hat{u} \in \arg\min_{u \in \mathbb{R}^{\omega}} \frac{1}{2} \left\| y - D^T u \right\|_{\frac{2}{2}} \text{ subject to } \|u\|_{\infty} \leq \lambda.(\#eq:dual) \tag{8}$$

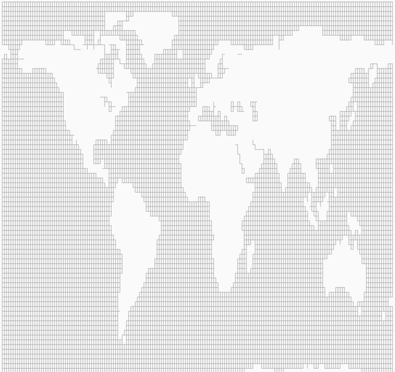
- The primal and dual solutions, $\hat{\beta}$ and $\hat{u}$ are related by:

$$\hat{\beta} = y - D^T \hat{u}.(\#eq:dual-relate) \tag{9}$$

- For general X and D with exploitable structure (as in our case), specialized implementations exist

Graph structure

- We can use a graph as input in the fusedlasso function
- We created a grid and deleted all nodes that were land regions
- This induced subgraphs



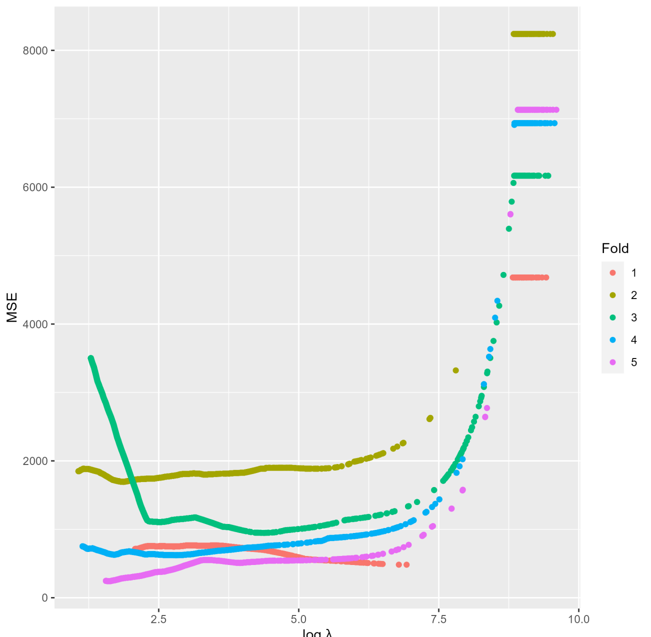
Graph structure and implications

- Results showed that removing the sub-graphs improved performance, although some of the regions were included in the final lasso models
- If we don't remove the clusters and also add sparsity (i.e $\gamma > 0$) the clusters dominate the results even more
- Possible explanations: Sub-graphs are less penalized, because they have fewer edges.
- Removing the clusters improved results more than f.e standardization

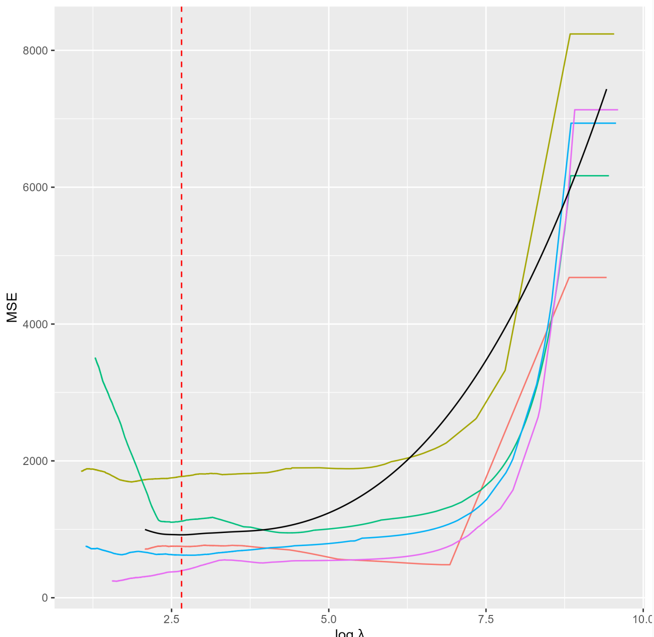
Fused lasso settings

- The considered fused lasso settings are: Fused lasso with clusters, fused lasso without clusters, fused lasso without clusters and sparsity (gamma: 0.01, 0.05, 0.1)
- Fused lasso without clusters and no sparsity showed best results

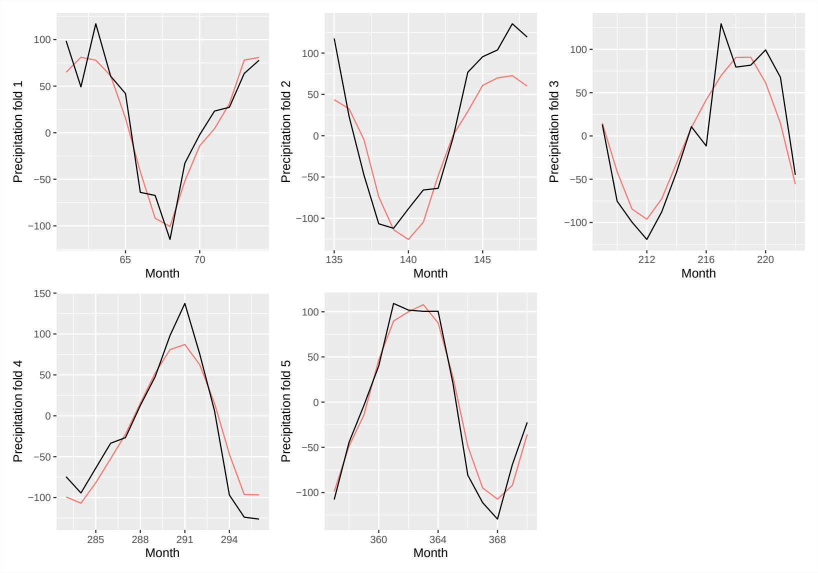
Fused lasso results, clusters removed



Fused lasso results, clusters removed



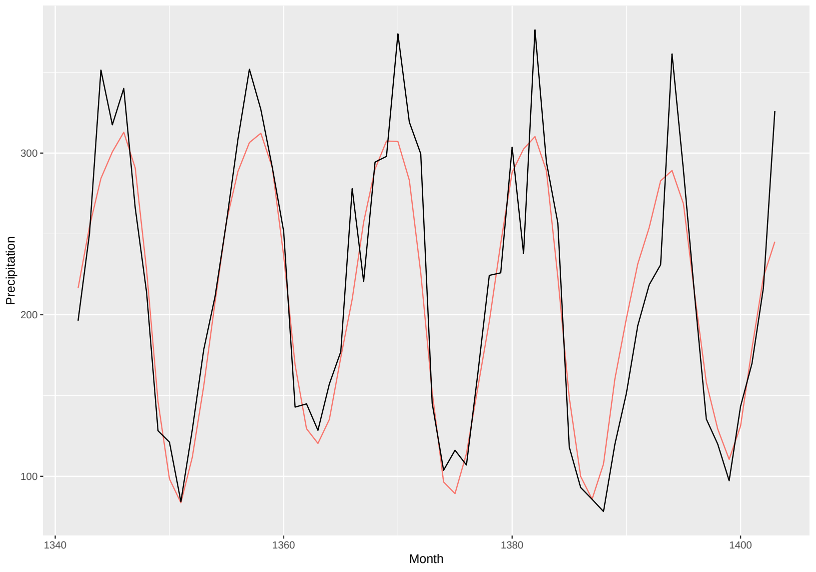
Prediction plots



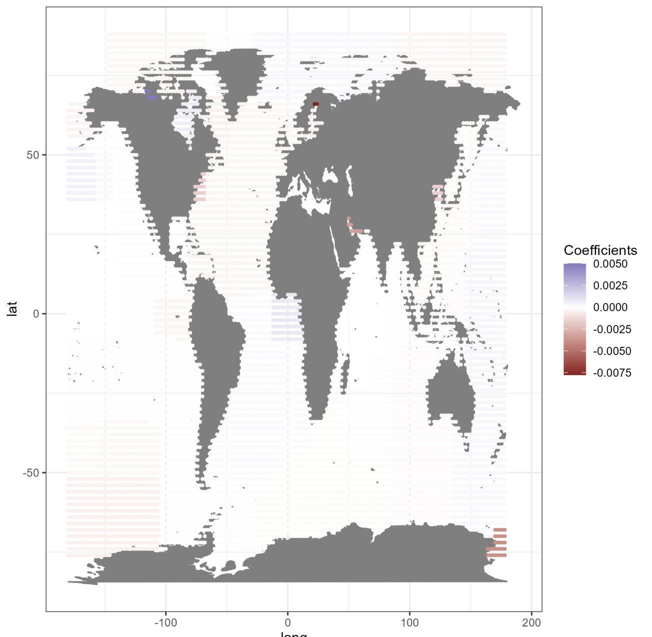
Prediction plot

The predictions inside the folds are very similar to lasso without standardization, the same holds for the predictions from the full model, but the MSE improves here.

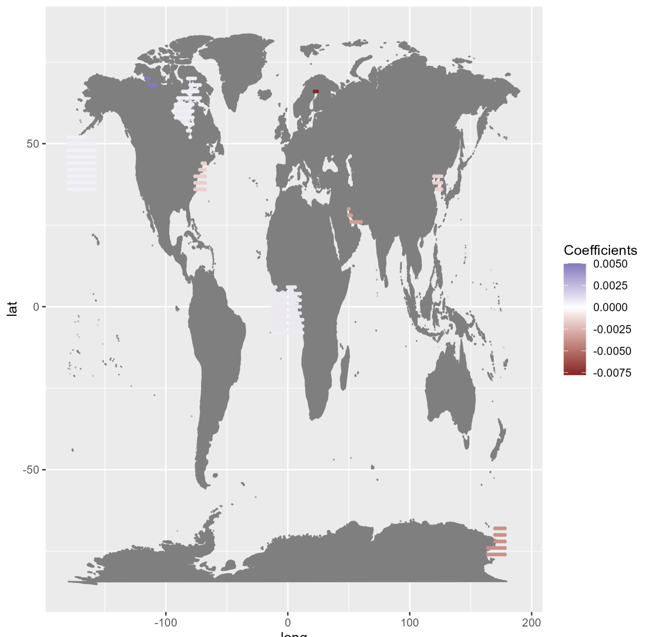
Full predictions



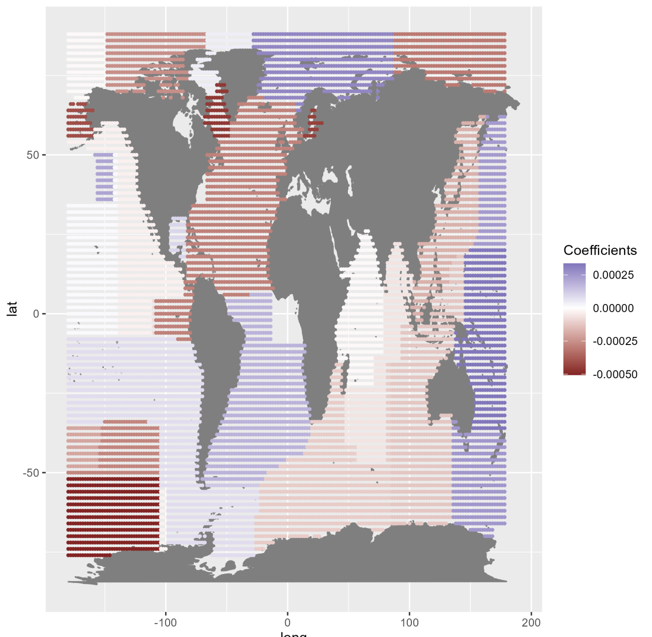
Coefficient plot



Coefficient plot, highest values only



Coefficient plot, lowest values only



Fused evaluation (maybe explain this when showing resutls)

- Generally same setting as for lasso, 5 folds with train and test, choose λ_{min}, refit with λ_{min}, get MSE on hold-out test set.
- But for the fused lasso we can not define the λ vector beforehand.
- λ-path is found by dual path algorithm and the range of the paths can vary a lot!
- So to find λ_{min} we search over the common range of all folds and interpolate to lines
- λ_{min} is then the λ that minimize MSE over all λ of that common range

Graph structure

Fused results

Summary

Important test

Important test

Ciemer, Catrin, Lars Rehm, Juergen Kurths, Reik V Donner, Ricarda Winkelmann, and Niklas Boers. 2020. "An Early-Warning Indicator for Amazon Droughts Exclusively Based on Tropical Atlantic Sea Surface Temperatures." *Environmental Research Letters* 15 (9): 094087.