CONFESS

Consistent representation of temporal variations of boundary forcings in reanalyses and seasonal forecasts

Report on the definition and performance of an empirical model for biomass burning emissions, benchmarked against climatology

Lead Authors: Alejandro Jiménez, Egor Tiavlovsky, Giulia Carella, Carlos Gómez, Etienne Tourigny, Markus Donat and Pablo Ortega

www.confess-h2020.eu

Co-ordinated by ECMWF

CONFESS

Consistent representation of temporal variations of boundary forcings in reanalyses and seasonal forecasts

D2.2: Report on the definition and performance of an empirical model for biomass burning emissions, benchmarked against climatology

Author(s): Alejandro Jiménez (BSC), Egor Tiavlovsky (BSC), Giulia Carella (BSC), Carlos Gómez (BSC), Etienne Tourigny (BSC), Markus Donat (BSC), and Pablo Ortega (BSC)

Dissemination Level: Public

Date: 27/10/2022

Version: 1.0

Contractual Delivery Date: 31/10/2022

Work Package/ Task: WP2/ T2.2

Document Owner: Organisation

Contributors: BSC

Status: For Review

ECMWF METEO FRANCE CNRM BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación ISAC

CONFESS

Consistent representation of temporal variations of boundary forcings in reanalyses and seasonal forecasts

Research and Innovation Action (RIA)
H2020- LC-SPACE-18-EO-2020 Copernicus evolution: Research activities in support of the evolution of the Copernicus services - Copernicus Climate Change Service (C3S

Project Coordinator: Dr Magdalena Alonso Balmaseda (ECMWF)
Project Start Date: 01/11/2020
Project Duration: 36 months

Published by the CONFESS Consortium

Contact:
ECMWF, Shinfield Park, Reading, RG2 9AX, United Kingdom
Magdalena.Balmaseda@ecmwf.int

The CONFESS project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004156.

Contents

Figures

Figure 9: Calibration scores for multivariate linear regression model with LASSO regularization using CIs (lagged by 1 to 12 months) and land cover as predictors. In all cases, the scores are averaged for the 6 different holdouts.
Figure 10: Test scores for multivariate linear regression model with LASSO regularization using CIs (lagged by 1 to 12 months) and land cover as predictors. In all cases, the scores are averaged for the 6 different holdouts.

Figure 11: Calibration scores for multivariate linear regression model with LASSO regularization using CIs (lagged by 1 to 12 months) and 1-month lagged emissions as predictors. In all cases, the scores are averaged for the 6 different holdouts.

Figure 12: Test scores for multivariate linear regression model with LASSO regularization using CIs (lagged by 1 to 12 months) and 1-month lagged emissions as predictors. In all cases, the scores are averaged for the 6 different holdouts.

Figure 13: Calibration scores for multivariate linear regression model with LASSO regularization using CIs (lagged by 4 to 12 months) and 4-month lagged emissions as predictors. In all cases, the scores are averaged for the 6 different holdouts.

Figure 14: Test scores for multivariate linear regression model with LASSO regularization using CIs (lagged by 4 to 12 months) and 4-month lagged emissions as predictors. In all cases, the scores are averaged for the 6 different holdouts.

Figure 15: Calibration scores for Random Forest regression model using CIs (lagged by 1 to 12 months) and 1-month lagged emissions as predictors. In all cases, the scores are averaged for the 6 different holdouts.

Figure 16: Test scores for Random Forest regression model using CIs (lagged by 1 to 12 months) and 1-month lagged emissions as predictors. In all cases, the scores are averaged for the 6 different holdouts.

Figure 17: Calibration scores for Random Forest regression model using CIs (lagged by 4 to 12 months) and 4-month lagged emissions as predictors. In all cases, the scores are averaged for the 6 different holdouts.

Figure 18: Test scores for Random Forest regression model using CIs (lagged by 4 to 12 months) and 4-month lagged emissions as predictors. In all cases, the scores are averaged for the 6 different holdouts.

Figure 19: Calibration scores for Random Forest regression model (opt. parameters) using CIs (lagged by 1 to 12 months) and 1-month lagged emissions as predictors. In all cases, the scores are averaged for the 6 different holdouts.

Figure 20: Test scores for Random Forest regression model (opt. parameters) using CIs (lagged by 1 to 12 months) and 1-month lagged emissions as predictors. In all cases, the scores are averaged for the 6 different holdouts.

Figure 21: Calibration scores for Random Forest regression model (opt. parameters) using CIs (lagged by 4 to 12 months) and 4-month lagged emissions as predictors. In all cases, the scores are averaged for the 6 different holdouts.

Figure 22: Test scores for Random Forest regression model (opt. parameters) using CIs (lagged by 4 to 12 months) and 4-month lagged emissions as predictors. In all cases, the scores are averaged for the 6 different holdouts.

Figure 23: Predictor importance, represented as the absolute value of the regression coefficients (for the linear regression with LASSO regularisation; left) or the feature importance (for the optimised Random Forest regression; right) for each predictor, considering the lag for which it is maximum. All values are derived for the regressions using CIs (lagged by 1 to 12 months) and 1-month lagged emissions as predictors.

Tables

Table 1: List of Climate indices

Table 2: Global average of scores and percentage of FCR for which the model outperforms the climatology benchmark in each regression method performed. Prior to computing the global averages, scores for the individual FCRs are averaged for the 4 seasons and the 6 holdout period selections. Values between brackets represent the same score averages/percentages but weighting each FCR contribution by its associated total accumulated emissions in the study period.

1 Executive Summary

This deliverable covers the different steps that were performed to develop an empirical model for predicting biomass emissions at the regional level, conceived to pave the way to the future implementation of a capability that can anticipate the occurrence of large biomass burning events and thus allow their inclusion in operational forecasts.

The model has been built using biomass burning emissions from the Global Fire Assimilation System (GFAS), which were aggregated over Fire Cohesive Regions (FCRs), and several well known climate and meteorological predictors. The potential predictive role of other local features, like the land cover type or the occurrence of previous burning emission events has also been considered. Different linear and non-linear regression methods have been tested in cross-validation mode and optimised to minimise overfitting. Several combinations of predictors, and forecast horizons have been considered, as well.

The best performing model overall is a multivariate linear regression with LASSO regularisation that includes climate indices and lagged emissions as predictors, which outperforms the climatology benchmark in more than half of the FCRs when predicting emissions 1-month in advance, and in more than a third of the FCRs when predicting emissions 4-months in advance. From all the predictors considered, the preceding emissions are the ones showing, by far, the largest predictive power worldwide.

A final account of the main lessons learned throughout this study, and a list of follow-up ideas for further improving the empirical model are provided at the end of the deliverable.

2 Introduction

2.1 Background

One of the main scientific goals of WP2 is to further our understanding on the predictive role that biomass burning emissions play on the climate system. The ultimate ambition is to build capabilities that can leverage their predictive capacity upon the climate and pave the way for their integration in the operational systems of the European Centre for Medium-Range Weather Forecasts (ECMWF).

All the activities building towards the fulfilment of these ambitious goals are performed in Task 2.2 of CONFESS. These involve the creation of an updated climatology with the latest observed biomass burning emissions, introducing a new feature in the ECMWF seasonal prediction system to use time-varying observed emissions throughout the forecasts, as well as performing two different sets of retrospective seasonal predictions to test the sensitivity of the forecasts to the emissions employed (climatological vs time-varying), paying particular attention to selected case studies of the most recent wildfire events. In real-time forecasts, since the evolution of biomass burning emissions is in essence unknown once the new forecast is launched, a climatology is generally applied. However, if we were able to anticipate, to some extent, how biomass burning emissions would evolve several months in advance, then operational predictions could, at least, partly benefit from their predictive capacity. Several recent studies point to a potential preconditioning role of different climate drivers on the occurrence of wildfires (Chen et al. 2011, 2016, 2020; Fernandes et al., 2011; Coscarelli et al. 2021), opening the door to their predictability, and in turn, the predictability of the associated biomass burning emissions. The last activity contemplated in Task 2.2 is therefore to build an empirical model to directly predict the monthly changes in biomass burning emissions, and to benchmark it against a climatology of emissions (i.e. the data currently used operationally).

2.1.1 Objectives of this deliverable

1) To build an empirical predictive model of biomass burning emissions that can clearly outperform the benchmark climatology forecast
2) To test different empirical models and methodological choices to explore their relative weaknesses and strengths
3) To investigate the predictive role of different types of predictors, both at the global and the regional level

2.1.2 Work performed in this deliverable

This deliverable documents all the different steps and methodological decisions undertaken to develop the empirical model (section 3), presents the major results with a particular focus on the aspects in which the model outperforms the climatology (section 4) and concludes with the main lessons learned and a list of ideas to further improve the model (section 5).

2.1.3 Deviations and countermeasures

This work was started by Egor Tiavlovsky, who developed the methodological framework, but left the BSC before completing the analyses. The work was resumed a few months later by Alejandro Jiménez, who finished the analyses and has led the writing of the deliverable. Because of the associated delays, the deadline of the deliverable was postponed with the approval of the project officer from the 30th of April 2022 to the 31st of October 2022.

3 Methodology

Generally, empirical models based on regression methods aim to estimate a relationship between the predictand (also known as dependent variable) and a series of variables that are deemed to have predictive skill over the predictand.

In this particular case, the predictands are biomass burning emissions from different geographical areas, taken from the Global Fire Assimilation System (GFAS) and then preprocessed to represent monthly deviations from climatology inside the Fire Cohesive Regions defined by Chen et al. (2020), as explained below in section 3.2.

Different combinations of predictors have been considered. In all cases we include a selection of Climate Indices (CI), which contain ten different oceanic indices and the North Atlantic Oscillation (NAO) index. Land cover data has been considered for some cases, as well as lagged values of the predictand. Their selection is justified in section 3.3.

3.1 Regression models applied

A handful of regression models of increasing complexity have been tested in an attempt to optimise the prediction skill achievable using the available predictors: namely, a multivariate linear regression, a multivariate linear regression with a LASSO regularization, and a Random Forest regression. The models have been implemented using the scikit-learn package (Pedregosa, 2011).

3.1.1 Multivariate linear regression

First, a simple multivariate linear regression was tested. In a linear type of regression, the predictand is considered as a linear function of the *p* predictors (x_1,...,x_p), in which each predictor is multiplied by a linear coefficient (w_1,...,w_p). An intercept coefficient is also included (w_0). The predicted $\hat{y}$ is thus determined by the equation:

$$\hat{y}(w, x) = w_0 + w_1 x_1 + \ldots + w_p x_p$$

The objective is to get the most accurate $\hat{y}$ possible, so the model is trained to find the coefficients that minimize the residual sum of squares between observed predictand (y) and its prediction by the regression model $\hat{y}$. This is done with the cost function:

$$\min_w ||Xw - y||_2^2$$

where y is the observed data vector for the predictand, X is the matrix of independent variables, and w is the vector of coefficients.

For this model to work properly, there should be a previous careful selection of the independent variables. If not, problems can arise if there is collinearity between predictors. When this happens, simple linear models can become too sensitive to outliers or errors, obtaining very different

parameters each time they are trained. It is said that the model has a high variance when this happens.

Additionally, without a proper selection of predictors to ensure that they are physically sound, an overfitting problem can occur. Overfitting happens when the model is fitted so closely to the training data that it affects its generalisation to make predictions with holdout data, leading to poor skill values when applied to independent testing datasets. The chances of overfitting will increase in situations in which there is a lack of training samples to properly capture the full mechanism, predictors chosen for the regression have collinearity between them, or some predictors do not have real predictive capacity over the predictand.

3.1.2 Lasso regression

For this project, a considerably large list of predictors was available. Some of them are oceanic indices from several parts of the globe, which are known to have covariance between each other. To solve the multicollinearity that causes problems in linear regressions, some methods exist to reduce the possible variance of the model. One of them is adding a LASSO (Least Absolute Shrinkage and Selection Operator) regularization to the multivariate linear regression. The LASSO penalty works by adding an extra term to the cost function of the linear regression method:

$$\min_{w} \frac{1}{2n_{\text{samples}}} ||Xw - y||_2^2 + \alpha ||w||_1$$

In our study, the regularization coefficient (alpha) is chosen with a 5-fold cross-validation.

The objective of this method is to ensure the selection of a reduced subset that only includes relevant lagged predictors. To this end the regularization method gives very little or zero weights to the predictors that do not show any covariance with the predictand. By choosing the “best” predictors for the regression (Friedman *et al*., 2010), the LASSO method aims to solve the variance and overfitting problems that usually arise in linear regression methods.

3.1.3 Random Forest

As an additional method to compare with the former two, a Random Forest regression has been trained with the same predictors. This non-linear regression method combines the predictions of several decision trees (weaker estimators) to enhance the robustness of the model.

Each decision tree is created gradually, making decisions step by step with a set of if-then-else decision rules. The random forest regression uses an ensemble of decision trees in which each tree is built from a bootstrap sample (smaller samples drawn with replacement), and averages the results of all the trees to achieve a more accurate and stable prediction (see schematic in Figure 1), which also helps reducing the total variance of the model (Breiman, 2001).

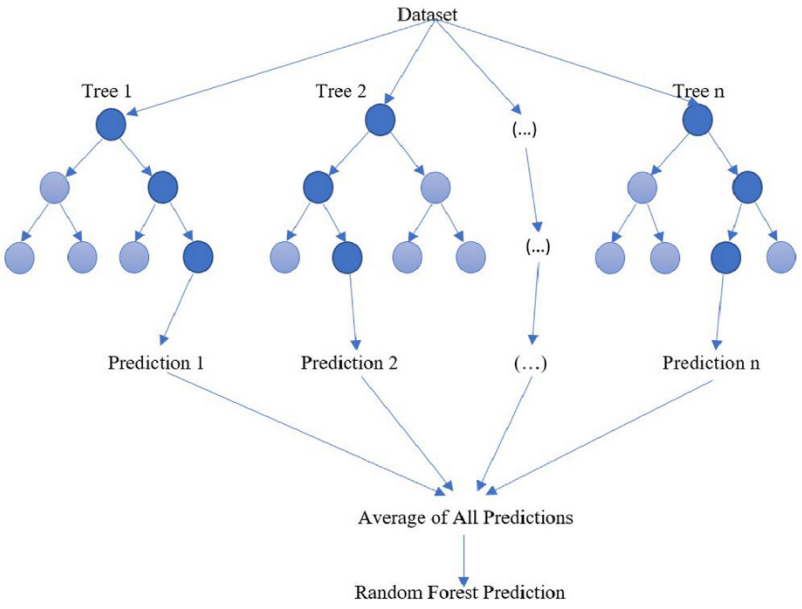

Figure 1: Schematic from a random forest regression (Figure from Uyanık et al 2022)

3.2 Spatial aggregation of burned emission data (predictand pre-processing)

Global monthly time series of GFAS fire emissions data from 2003 to 2020 at a 0.25°x0.25° resolution had been initially considered as our predictand. This dataset, however, came with a problem: the ubiquity of zeros (months without fires) in time series. To deal with this, we use emissions that have been first re-gridded at a 1°x1° resolution, and spatially aggregated data over Fire Cohesive Regions or FCRs. These regions are defined to ensure that their associated fire emission time series are suitable for the application of linear regression methods, and were kindly shared with us by the producers (Chen et al., 2020).

The size of an FCR can vary, with the options being 1°x1°, 2°x2°, 4°x4° and 8°x8°. Following the criteria in Chen et al. (2020) a 1°x1° grid cell is defined as an individual FCR if:

1. Total fire emissions on the area are equal or above a minimum threshold of 1.33 Tg C/yr.
2. More than 90% of the years of the time series contain non-zero emissions.
3. The coefficient of variation of the timeseries (i.e. the ratio of standard deviation to the mean) is less than 2, so that the variability is not controlled by sparse (and anomalously large) data samples.

If a region does not fulfil the three points, the fire emissions of neighbouring grid points that did not satisfy the criteria are aggregated to consider a 2°x2° FCR, and the same above conditions are checked. The process continues until the 8°x8° FCRs are defined. The remaining areas that failed to meet the three criteria (mostly located in deserts and polar regions), were assumed to be fire-free and excluded from the forecasts. Resulting FCRs are then subdivided with the borders of different

countries, to reflect potentially important differences across countries in terms of fire management policies and resources.

3.3 Predictor selection

3.3.1 Climate Indices (CIs)

Lagged monthly ocean climate indices have been proven to have significant predictive skill for past fire emissions in several FCRs (Chen et al., 2016). For this reason, an updated selection of the major ocean CI time series considered in Chen et al. (2016) has been considered for our regression models. In addition, an atmospheric index, the North Atlantic Oscillation (NAO), has been added to the list, as it is known to influence European precipitation, which could potentially precondition several months before the local risk of wildfires. We have considered their predictive value on fire emissions 1 to 12 months ahead. The final list of observed Climate Indices considered, including the NAO index, is shown in Table 1. They are all obtained either from the WMO website (https://climexp.knmi.nl/selectindex) or the NOAA website (Climate Indices: Monthly Atmospheric and Ocean Time Series: NOAA Physical Sciences Laboratory).

Table 1: List of Climate indices

CI	Data	Method
Tropical North Atlantic (**TNA**)	HadISSTv1	Mean sea surface temperature anomaly (SSTA) in the box 15°W - 57.5°W, 5.5°N - 23.5°N, climatology: 1971-2000
Tropical South Atlantic (**TSA**)	HadISSTv1	Mean SSTA in the box 10°E - 30°W, 20°S - 0, climatology: 1971-2000
South Western Indian Ocean (**SWIO**)	HadISSTv1	Mean SSTA in the box 31°E - 45°E, 32°S - 25°S, climatology: 1971-2000
NIÑO 1.2	HadISSTv1	Mean SSTA in the box 90°W - 80°W, 5°S - 5°N, climatology: 1971-2000
NIÑO 4	HadISSTv1	Mean SSTA in the box 160°E - 150°W, 5°S - 5°N, climatology: 1971-2000
NIÑO 3	HadISSTv1	Mean SSTA in the box 150°W - 90°W, 5°S - 5°N, climatology: 1971-2000
Pacific Decadal Oscillation (**PDO**)	1900-81: UKMO SST 1982-2001: OIv1SST 2002-present: OIv2SST	Leading EOF of mean November through March SSTA in the box 100°E-70°W, 0- 20°N
Western Tropical Indian Ocean (**WTIO**)	HadISSTv1	Mean SSTA in the box 50°W - 70°W, -10°S - 10°N, climatology: 1971-2000
SouthEast Indian Ocean (**SEIO**)	HadISSTv1	Mean SSTA in the box 90°W - 110°W, -10°S - 0, climatology: 1971-2000
Atlantic Multidecadal Oscillation (**AMO**)	HadISSTv1	SST 0-60°N, 0-80°W minus SST 60°S-60°N, climatology: 1971-2000
North Atlantic Oscillation (**NAO**)	NCEP/NCAR CDAS	First Empirical Orthogonal Function (EOF) of North Atlantic Sea Level Pressure

The data of these indices has been obtained from 2002 to 2020. A matrix of values for these indices has been created for a 2003-2020 monthly timeline. For each CI, 12 column vectors have been created. Each vector represents the value of the given index for lags from 1 to 12 months.

3.3.2 Preconditioning role of land surface

The local risk of wildfires (and subsequent emissions) can be largely preconditioned by the local land cover type. For that reason, we have considered the inclusion of global timeseries of land cover data as a complementary predictive source to the CI indices. Gridded land cover classification maps from 2002 to present from the Copernicus database were used to this end. The data have been regridded to a 1°x1° resolution and then aggregated to match the FCRs. For each region, land cover data has been encoded as a multidimensional vector representing the fraction of gridpoint belonging to each of the various land cover categories. The vector is then appended to the rest of predictors for each FCR. There are originally 30 different categories of land cover class defined, based on the UN Land Cover Classification System (LCCS). The global spatial distribution of land cover types is illustrated in Figure 2 for an example year (2008).

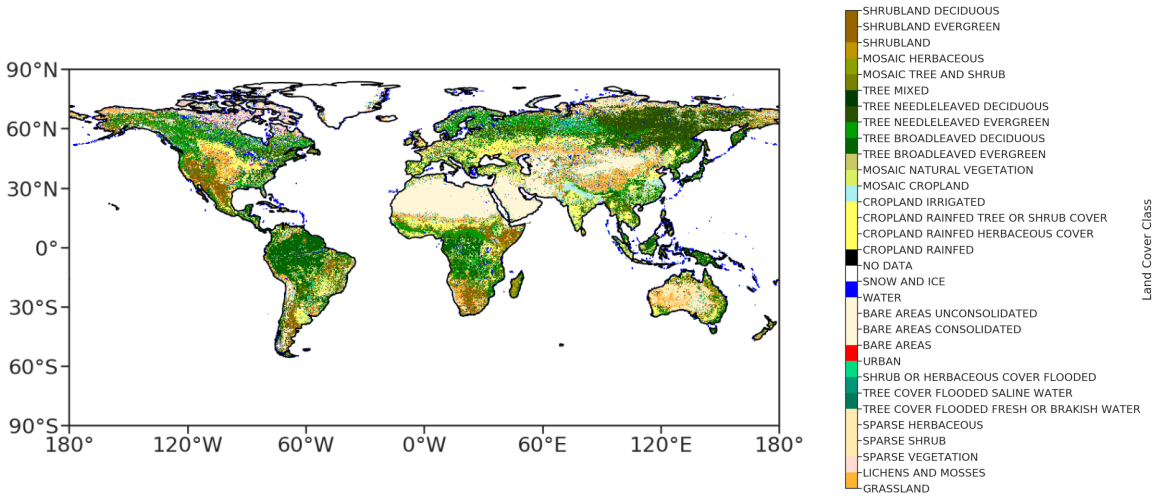

Figure 2: Spatial distribution of land cover data for the year 2008

3.3.3 Predictive potential of biomass burning emissions

Biomass burning emissions themselves also have some predictive potential on the future emissions, as evidenced by the autocorrelation maps in Figure 3 for 1 month and 4 month lags. The autocorrelation values at lag 1 (which illustrate the predictive potential of emissions one month ahead) are positive everywhere, with values that in many regions (like South America or Central Africa) reach well over, e.g., 0.5. Autocorrelation at lag 4 (illustrating predictive potential 4 months ahead) tends to be smaller in magnitude, but also important in regions where large emissions occur like central and south Africa, showing negative values that can locally surpass the -0.4 threshold. These negative correlations could be explained by the fact that several months after important biomass burning events occur, the amount of fuel available for subsequent events gets substantially reduced.

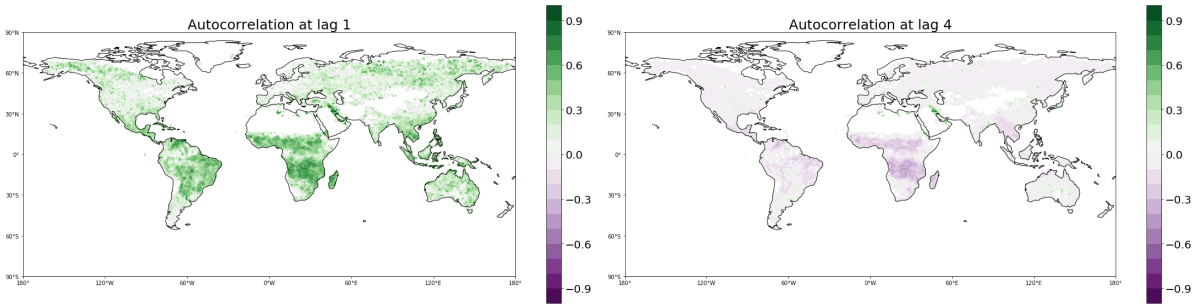

Figure 3: Temporal autocorrelation of biomass burning emissions per FCR based on de-seasonalised monthly means (at lag 1 on the left, and lag 4 on the right). The regions that did not meet the requirements to be qualified as FCRs were excluded from the analysis and appear colored in white.

3.4 Other methodological aspects

Due to the short span of fire emissions observations, which only cover 18 years, the regressions are built on monthly data, to thus have a longer sample to train and test the models. This implies that to be able to fully benefit from the temporal sample of independent 216 months, the models need to be trained with regression coefficients that are the same throughout the year (i.e. seasonally independent) for each predictor. This thus comes at the expense of diluting the predictive potential of predictors that have a marked seasonality (e.g. the NAO), a problem that is largely compensated by the much longer sample to train the model, which is critical to identify meaningful predictive relationships.

We removed the annual cycle that is present in monthly data due to seasonal variations, which would provide trivial predictability, and computed standardized anomalies for all predictors and emissions data, to in this way build our regression models with non-dimensional metrics. For consistency, the CI predictors that were directly provided as monthly anomalies have been de-seasonalized again to have anomalies defined with respect to the years for which emissions data are available. The model is trained with these pre-processed predictors, thus constructed on the anomaly space.

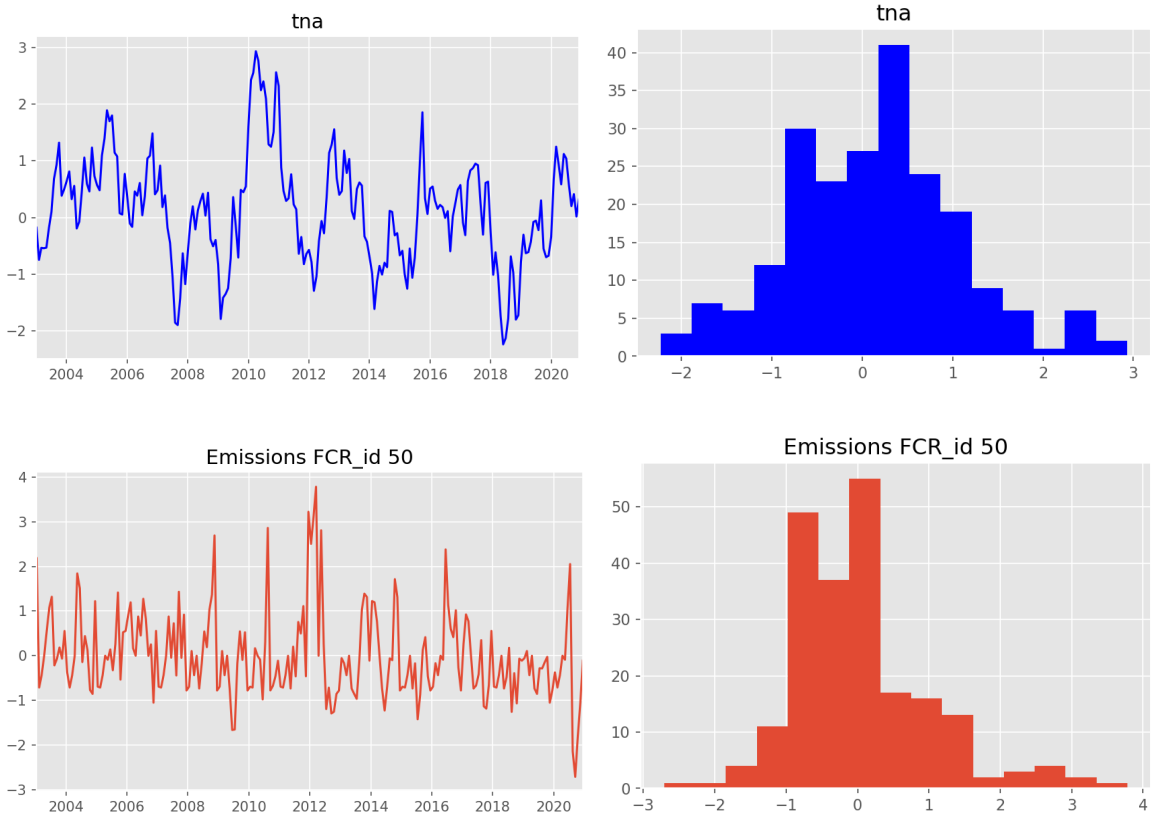

Figure 4: Histograms and time series for an example of processed predictors and predictands. The left column represents the standardised monthly time series of the TNA climate index at the top, and the biomass burning emissions at the FCR with id number 50 at the bottom. The right column shows a histogram of all the monthly values for the corresponding time series.

By construction, the time series of the resultant pre-processed predictors show deviations from the climatological seasonal cycle, which have identifiable intra-annual and also interannual variations, as illustrated in the left column of Figure 4 for one example predictor, in this case the TNA. Similar features are also seen in the pre-processed predictand for a given FCR (Figure 4, left column, second row). Histograms of the final processed feature predictors and predictand (examples in the right column of Figure 4) show distributions that can be roughly approximated by gaussian functions, and are therefore suitable for the application of linear regression techniques.

3.5 Evaluation of the predictive model

3.5.1 Evaluation metrics

By working in anomaly space, the model has been trained to estimate the deviations of fire emissions from their monthly climatology. So in absence of any significant predictive skill, the model could default to climatology. To evaluate the added value of our regression models, we compare with a simple model that assumes the climatological value, using a score that measures the differences between mean squared errors (MSE) of both models:

$$\text{score} = \mathbf{MSE}(\hat{Y}_{clim}, Y) - \mathbf{MSE}(\hat{Y}_{reg}, Y)$$

where $\hat{\mathbf{Y}}_{reg}$ is the prediction from the regression model, $\mathbf{Y}$ is the observed variable to predict, and $\hat{\mathbf{Y}}_{clim}$ is a prediction that assumes no deviation from the climatological value.

The first term at the right-hand side of the equality represents the error committed by the climatology values. The second term is the error committed by the regression model. So, by construction, a positive score will mean that the model performs better than climatology, and the opposite will happen if it is negative (in which case the final model could default to climatology).

3.5.2 Calibration and validation procedure

The model has been evaluated according to its performance against climatology on a holdout test set (which has been rotated and averaged using a 6-fold cross-validation). From the full record of the 18 years for which the emission data is available (2003-2020), the following holdout sets have been randomly generated:

[2006, 2017, 2004]	[2005, 2003, 2019]	[2015, 2008, 2013]
[2007, 2011, 2012]	[2020, 2016, 2010]	[2018, 2009, 2014]

Within each FCR, the model is trained separately for each holdout (selecting different regression coefficients in each case), with data from the years that are not included in said set, and evaluated afterwards for the holdout years. Both training and testing scores from the six holdouts are averaged to get a single score in each region for the training and testing periods.

4 Results

The results of the training and testing scores for all the different models designed are shown for each FCR. The results are presented in order of increasing complexity in the regression protocol.

The first model that was tested is the simple multivariate linear regression model, considering only CIs as predictors (section 4.1) and focusing on the first forecast month, as it is the one for which the best model performance is expected. Clearly better overall results were obtained when adding the LASSO regularisation technique, which justified keeping the regularised model for the subsequent tests (section 4.2). After that, we tested the impact of adding new types of predictors. Annual land cover data was added and tested in section 4.3, and lagged emissions in section 4.4.

From the previous tests the best performing model is the LASSO regularized linear regression that used CIs and lagged emissions as predictors (section 4.4.1).

We additionally evaluated the performance of the same model but four months ahead, to explore its potential use as refined boundary conditions in actual seasonal predictions. In this case we thereby used 4-month lagged emissions (instead of 1-month), and 4 to 12-month lagged CIs as predictors. The two additional sections of results correspond to two different versions of the Random Forest regression model. The first was implemented with the default parameters of the statistical package considered (section 4.5.1), and the second considered an extra restriction which forced the maximum depth of the trees used for the regression to be equal or lower than 2 (section 4.5.2). For each version, the model was trained two times: one to make predictions 1-month ahead, and another to make predictions 4-months ahead, similarly to the last versions of the LASSO regularised model. The intention is to compare the performances for both methodologies, which are respectively based on a non-linear and a linear model, using the same predictor selection. Likewise, in a final section (4.6) we compare the metrics of predictor importance as identified by both methodologies, to illustrate which predictors matter the most in the different FCR regions.

4.1 First results from a basic multivariate linear regression model

The first model is built with a basic multivariate linear regression method that uses only CIs as predictors. The monthly results for the training and testing periods are grouped by seasons, and then averaged for the 6 different holdout selections. The results are shown as spatial maps of the score values for each FCR and season of the year. The colorbar is bounded to [-1,1] and centred so it shows greater performance than climatology in regions shown in red, and blue otherwise.

These first results obtained with the simple multivariate linear regression model show a clear example of overfitting. Training scores are very high for each season in almost all the FCRs (Figure 5). However, the testing scores are predominantly negative (Figure 6), consistently showing much lower performance than the climatological benchmark across all seasons and almost every FCR region. A somewhat expected problem of this linear model is that it is built with such a large number of predictors for the monthly fire emissions data that it can reproduce the predictand variability very closely by learning spurious relationships with the predictors. As a result, the model performs very poorly when tested outside of the training sample.

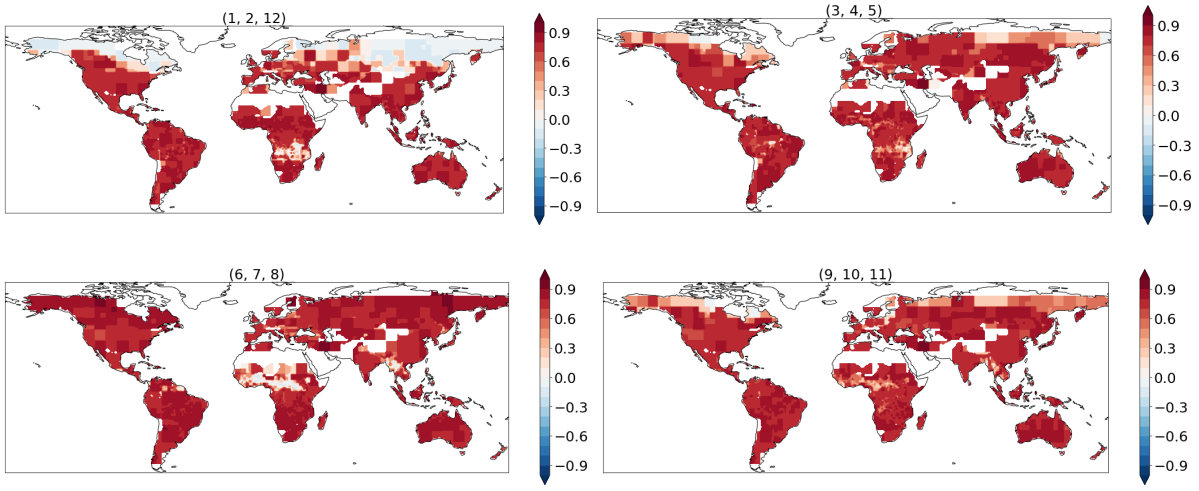

Figure 5: Calibration scores for the basic multivariate linear regression model using only CIs (lagged by 1 to 12 months) as predictors. Results are shown separately for winter (DJF; top left panel), spring (MAM; top right panel), summer (JJA; bottom left panel) and autumn (SON; bottom right panel). In all cases, the scores are averaged for the 6 different holdouts.

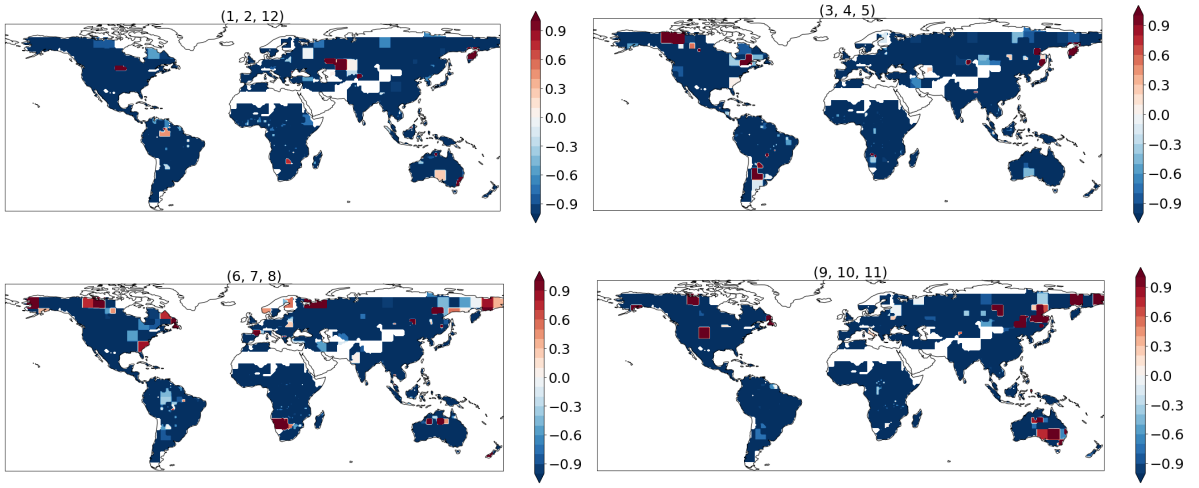

Figure 6: The same as in Figure 5 but for the test scores.

To deal with this issue, a LASSO regularisation that penalises the spurious predictors was implemented into the regression model.

4.2 Improving the multivariate linear regression model via LASSO regularisation

The LASSO regularization is first added to the previous reggression model using only the CIs as predictors. For the sake of simplicity, from now on, only results for boreal summer (JJA) and winter (DJF) are shown. In this and the subsequent empirical models the major conclusions drawn from the analysis of the winter/summer months are very similar to those from the analysis of the autumn/spring months.

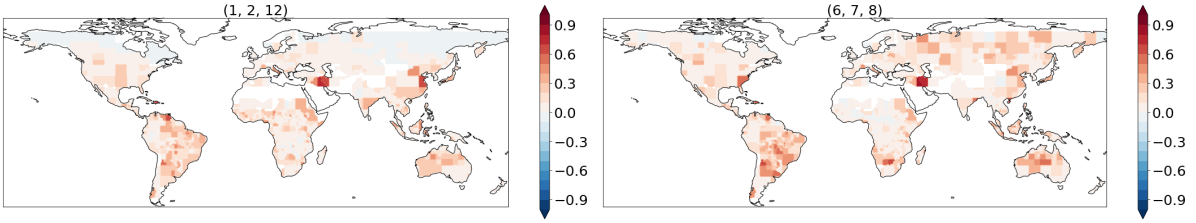

Figure 7: Calibration scores in winter (DJF; left) and summer (JJA; right) for multivariate linear regression model with LASSO regularization using only CIs (lagged by 1 to 12 months) as predictors. In all cases, the scores are averaged for the 6 different holdouts.

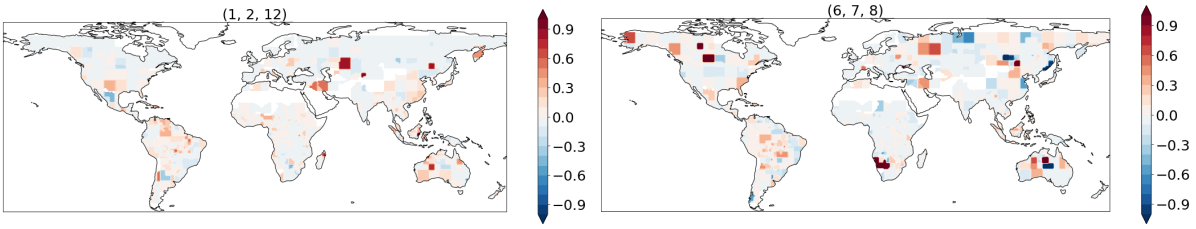

Figure 8: Test scores in winter (DJF; left) and summer (JJA; right) for multivariate linear regression model with LASSO regularization using only CIs (lagged by 1 to 12 months) as predictors. In all cases, the scores are averaged for the 6 different holdouts.

Adding the regularisation has a remarkable impact on the results. Training scores are notably lower compared to the non-regularised model (Figure 7), and, more importantly, testing scores are generally better for every season and almost every region (Figure 8). Some regions with certain characteristics can be recognized from the maps. At the very northern latitudes, the model is unable to find any predictive relationships in the winter months. This might be probably due to the lack of fire emissions in said regions (e.g. Siberia) during the cold season. There are substantially more regions that outperform the climatology benchmark than for the non-regularised case: for example, the FCRs located in the Amazon basin in South America show positive testing scores in both the boreal winter and summer months, which could derive from the large predictive role in local precipitation associated with ENSO. Somewhat surprisingly, some FCRs show substantially higher test scores than training scores (e.g. in Alaska in the summer months) which suggests that, despite the 6-fold cross-validation, the scores exhibit some level of uncertainty. It can also be seen that some regions still have poor scores indicating no added value over the climatology, like Central Africa during the boreal summer months, or Northern Asia in both seasons. Globally averaged results (Table 2, second row) confirm that the training scores are much lower than for the basic regression, as expected, and that there is a great improvement of the testing scores. Nevertheless, we also note that the total global scores are still negative, which is related to the fact that about 2 thirds of the FCRs do not outperform the climatology benchmark. The mean test score remains negative also when the global average is weighted by the accumulated emissions of each FCR (number in bracket in Table 2).

Table 2: Global average of scores and percentage of FCR for which the model outperforms the climatology benchmark in each regression method performed. Prior to computing the global averages, scores for the individual FCRs are averaged for the 4 seasons and the 6 holdout period selections. Values between brackets represent the same score averages/percentages but weighting each FCR contribution by its associated total accumulated emissions in the study period.

Regression method	Calibration scores global average (weighted average)	Test scores global average (weighted average)	% of positive FCRs in the test period (% of emissions)
Linear	0.6451 (0.6458)	-2.6265 (-2.4101)	1.52 % (2.15 %)
LASSO regularised	0.0876 (0.1006)	-0.0049 (-0.0089)	35.72 % (36.23%)
LASSO regularised + land cover	0.0876 (0.1006)	-0.0049 (-0.0089)	35.72 % (36.23%)
LASSO regularised + emissions (t-1)	0.1058 (0.1234)	0.0421 (0.0622)	53.40 % (57.05%)
LASSO regularised + emissions (t-4)	0.0761 (0.0872)	-0.0006 (0.0052)	35.86 % (36.02%)
Random Forest + emissions (t-1)	0.7190 (0.7173)	-0.0550 (-0.0181)	27.89 % (29.73%)
Random Forest + emissions (t-4)	0.7167 (0.7142)	-0.1165 (-0.0957)	18.11 % (19.22%)
Random Forest + emissions (t-1) (opt. parameters)	0.3287 (0.3384)	0.0022 (0.0275)	35.50 % (38.13%)
Random Forest + emissions (t-4) (opt. parameters)	0.3024 (0.3074)	-0.0448 (-0.0347)	24.34 % (25.66%)

4.3 Enhancing the model with the inclusion of time-varying land cover

Results are now shown for a new regression model that is exactly the same as the one developed in the previous section, but incorporating land cover data for the corresponding FCR as an additional predictor.

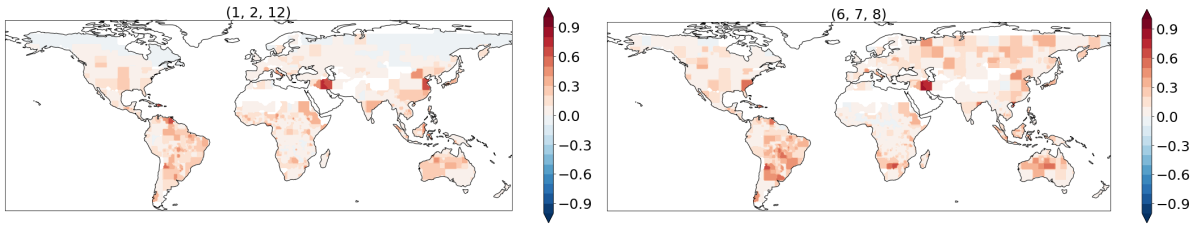

Figure 9: Calibration scores for multivariate linear regression model with LASSO regularization using CIs (lagged by 1 to 12 months) and land cover as predictors. In all cases, the scores are averaged for the 6 different holdouts.

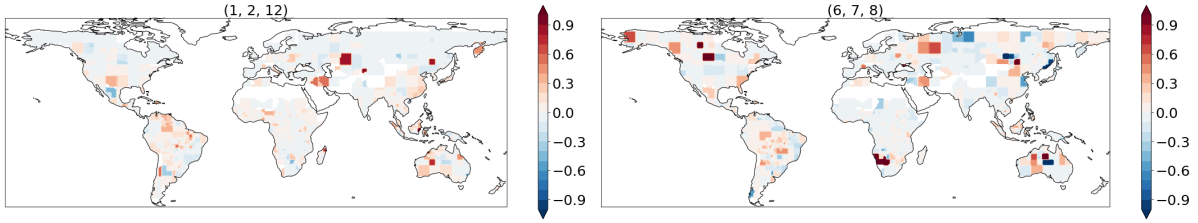

Figure 10: Test scores for multivariate linear regression model with LASSO regularization using CIs (lagged by 1 to 12 months) and land cover as predictors. In all cases, the scores are averaged for the 6 different holdouts.

Resulting maps are identical to the ones shown in the previous section (Figures 9 and 10 vs Figures 7 and 8), and the same happens when averaging the total scores (Table 2, second and third rows). These identically distributed values with respect to the previous ones can only mean that the LASSO regularisation gave zero coefficients to each of the land cover categories in every case, which implies that the regression method did not detect any meaningful linear relationship (and therefore predictive potential) between the land cover type variations and biomass burning emissions in the different FCRs.

Based on these findings we decided to dismiss the use of this type of predictors in the subsequent analyses, as they would add a computational burden to the regression, while adding no improvement to the predictions.

4.4 Enhancing the model with the inclusion of lagged emissions as predictors

Emissions data are added as predictor features for two different predictive horizons. This is an idealised exercise that explores to what extent these emissions could add predictive skill if their corresponding observations were available in real time to be included in the forecasts.

4.4.1 Lag 1 results

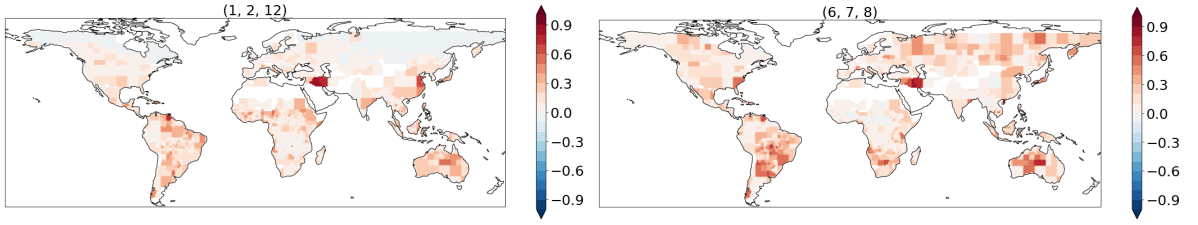

Figure 11: Calibration scores for multivariate linear regression model with LASSO regularization using CIs (lagged by 1 to 12 months) and 1-month lagged emissions as predictors. In all cases, the scores are averaged for the 6 different holdouts.

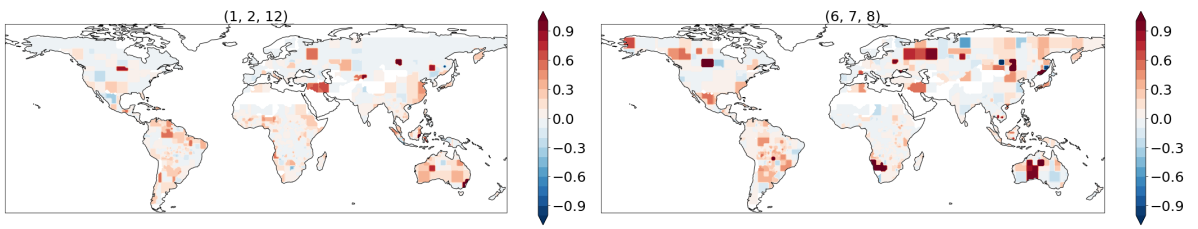

Figure 12: Test scores for multivariate linear regression model with LASSO regularization using CIs (lagged by 1 to 12 months) and 1-month lagged emissions as predictors. In all cases, the scores are averaged for the 6 different holdouts.

The addition of emissions data from the previous month to the predictor list results in a notable improvement with respect to the previous LASSO regularised models. It can be seen in the training score maps (Figure 11 vs Figure 7), but more notably in the test scores maps (Figure 12 vs Figure 8), that the regions where the model already performed better than climatology before adding the lagged emissions as predictors are maintained. And in addition, some other regions where the model previously underperformed the climatology benchmark now outperform it. This can be seen, for example, across the whole Eurasiancontinent in the months of boreal summer. There are also regions like South America and Australia where the regressions with lagged emissions show particularly better scores, two zones in which biomass burning emissions showed high temporal autocorrelation values at lag 1 (Figure 3). Interestingly, emissions with high autocorrelations at lag 1 do not guarantee good test scores, with the most clear example being the African continent in the boreal summer, for which only a few FCRs show positive score values.

The addition of lagged emissions data is also helpful in regions where none of the CIs data seemed to have a predictive effect on the biomassburning events. Performance improvement in these regions is attested by enhanced testing scores, which in a global average turn out to be greater than zero (Table 2), indicating that this model yields overall improvements with respect to the climatological benchmark.

4.4.2 Lag 4 results

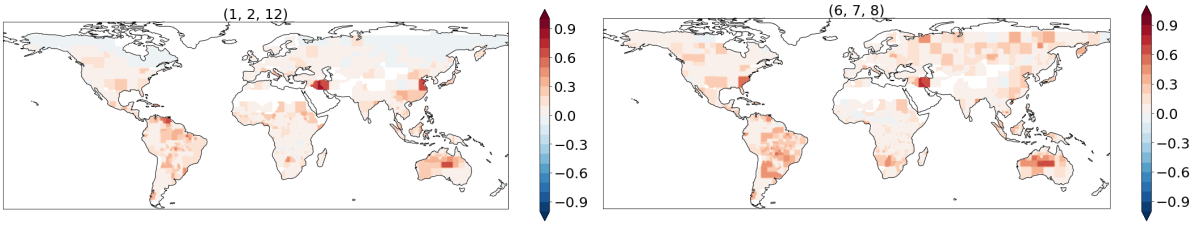

Figure 13: Calibration scores for multivariate linear regression model with LASSO regularization using CIs (lagged by 4 to 12 months) and 4-month lagged emissions as predictors. In all cases, the scores are averaged for the 6 different holdouts.

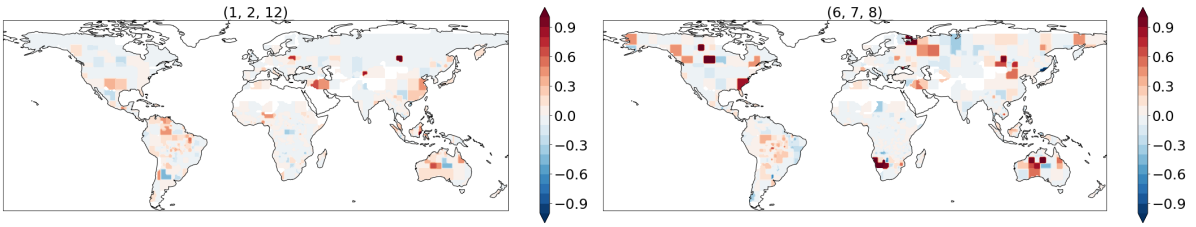

Figure 14: Test scores for multivariate linear regression model with LASSO regularization using CIs (lagged by 4 to 12 months) and 4-month lagged emissions as predictors. In all cases, the scores are averaged for the 6 different holdouts.

Results when we replicate the same method but with a set of predictors that precede the predictand with 4-month to 12-month lags show a much more limited predictive potential. Even if the training scores of the new model remain close to those of the regression model with 1-month lagged predictors (Figures 11 and 13), the testing scores are substantially lower in the new regression model (Figures 12 and 14), with only some sparse regions like the Amazon basin, Western Australia and Southeast Asia showing positive scores consistently throughout the seasons. Table 2 indeed shows that only 35% of all FCRs have positive test scores for the 4-month lagged predictions, which leads to

a negative global value when averaging all FCRs. Interestingly, the averaged test score that weights each FCR by its accumulated emissions is found to be positive, which indicates that the FCRs in which the regression outperforms the climatology contribute more actively to the total emissions.

4.5 Results from a predictive model based on the random forest technique

The set of predictors tested for the LASSO regularised linear regression in Section 4.4.1 showed the best performance when compared to the climatological reference. The same set of predictors is now used to build a Random Forest Regression model, to check whether this non-linear approach can improve the previous results. First, a regression with the default parameters of the random forest routine is performed.

4.5.1 Results with default random forest parameters

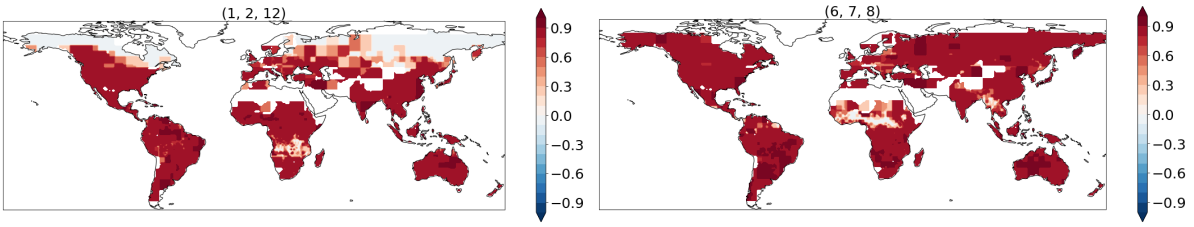

Figure 15: Calibration scores for Random Forest regression model using CIs (lagged by 1 to 12 months) and 1-month lagged emissions as predictors. In all cases, the scores are averaged for the 6 different holdouts.

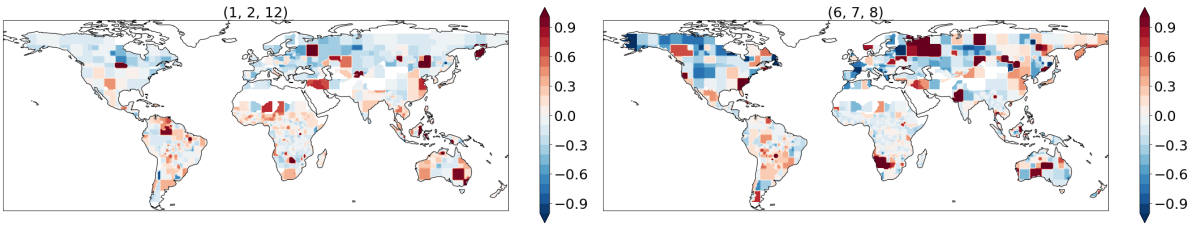

Figure 16: Test scores for Random Forest regression model using CIs (lagged by 1 to 12 months) and 1-month lagged emissions as predictors. In all cases, the scores are averaged for the 6 different holdouts.

When the model is trained with this default configuration to predict biomass burning emissions one month ahead, it consistently yields very good training scores (Figure 15), but it does not translate into good performance in the testing scores, which are more contrasted between regions (Figure 16). Compared with results from the linear models, the random forest test scores tend to be larger in magnitude and also noisier, with neighbouring FCRs seldom showing large and opposing scores. Also, the number of regions where the model performs better than the climatology is substantially lower than the number of regions in which it is outperformed by it, as shown in Table 2. In general, this results in very poor globally averaged testing scores. Results are very similar, but with slightly worse scores when predicting the biomass burning emissions 4 months ahead (Figures 17 and 18).

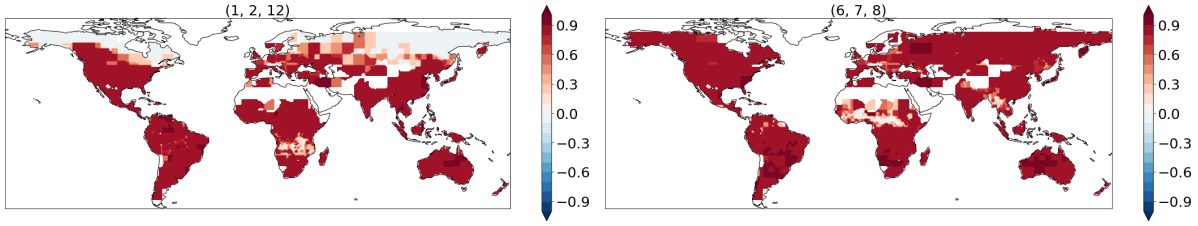

Figure 17: Calibration scores for Random Forest regression model using CIs (lagged by 4 to 12 months) and 4-month lagged emissions as predictors. In all cases, the scores are averaged for the 6 different holdouts.

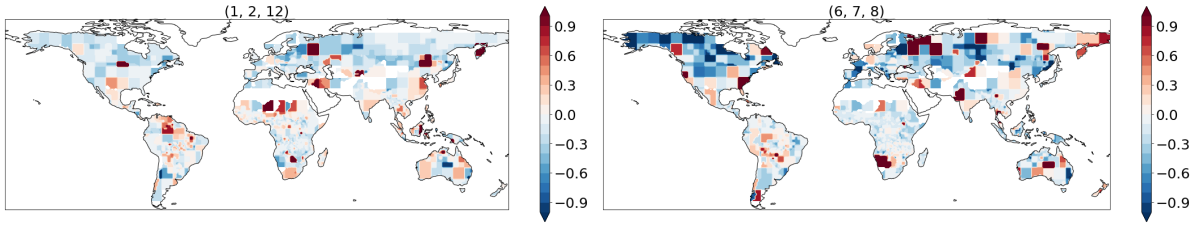

Figure 18: Test scores for Random Forest regression model using CIs (lagged by 4 to 12 months) and 4-month lagged emissions as predictors. In all cases, the scores are averaged for the 6 different holdouts.

4.5.2 Results with optimised random forest parameters

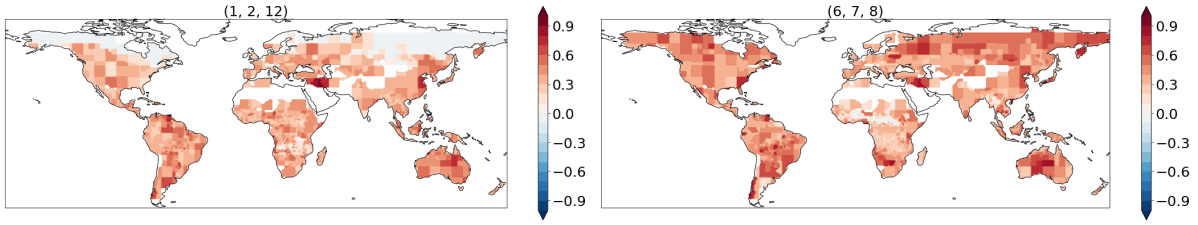

Figure 19: Calibration scores for Random Forest regression model (opt. parameters) using CIs (lagged by 1 to 12 months) and 1-month lagged emissions as predictors. In all cases, the scores are averaged for the 6 different holdouts.

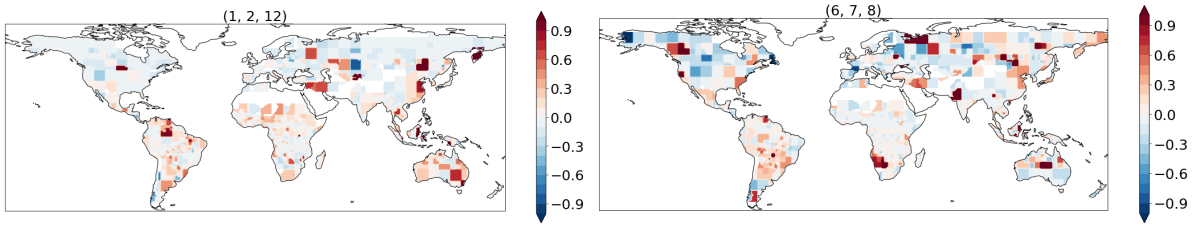

Figure 20: Test scores for Random Forest regression model (opt. parameters) using CIs (lagged by 1 to 12 months) and 1-month lagged emissions as predictors. In all cases, the scores are averaged for the 6 different holdouts.

A characteristic of the standard configuration for the random forest routine from the scikit-learn package is that its maximum depth parameter allows for very deep trees. This is known to frequently lead to overfitting, since deep trees always depict more complex predictive algorithms, which in some cases can cause trees to produce very weak testing results (contrasting with very high training scores), while the real connection mechanism between the predictors (or features) and the predictand is usually less intricate. The Random Forest method relies on averaging between all trees created to avoid this problem, but it can also be helpful to limit the maximum depth number for all

the trees in order to achieve a higher testing accuracy. This is particularly advantageous when the signal-to-noise ratio in the data is low as in our case (Zhou and Mentch, 2022).

Results for the same model but now forcing the maximum depth of the trees to be lower or equal than 2 show enhanced predictions of biomass burning emissions 1 month ahead in many regions (Figure 20), raising the globally averaged test score to a positive value (Table 2), which becomes even larger when the average is weighted by the accumulated FCR emissions. However, both the test score values and the percentage of regions where the regression model outperforms the climatological benchmark are considerably lower for this optimised random forest model than for the analogous case with the LASSO regularised model. The linear model is then currently preferred because it can capture predictive skill over more regions in the world.

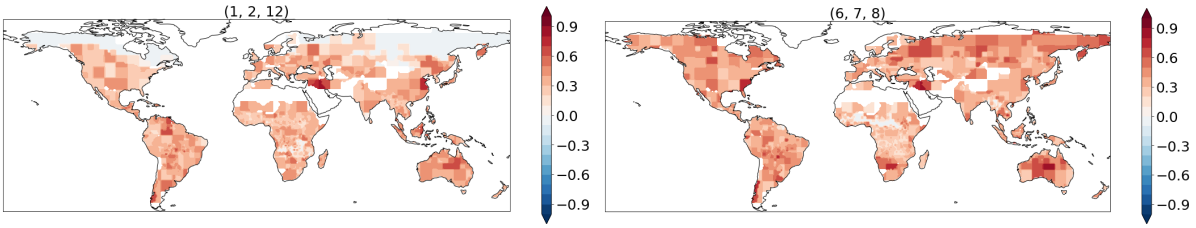

Figure 21: Calibration scores for Random Forest regression model (opt. parameters) using CIs (lagged by 4 to 12 months) and 4-month lagged emissions as predictors. In all cases, the scores are averaged for the 6 different holdouts.

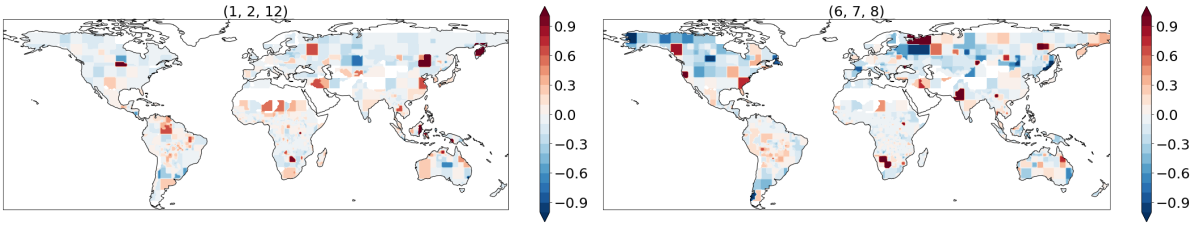

Figure 22: Test scores for Random Forest regression model (opt. parameters) using CIs (lagged by 4 to 12 months) and 4-month lagged emissions as predictors. In all cases, the scores are averaged for the 6 different holdouts.

Conclusions are once again very similar when performing predictions 4 months ahead (compare Figure 20 with Figure 22).

4.6 Assessing how the relative importance of the predictors varies regionally

To compare the predictive power of the different predictors, and whether it is consistent across methods, we now plot the predictor importance for the best performing models of both the linear regression with LASSO regularisation, and the optimised Random Forest regression (Figure 23). This importance corresponds to the absolute value of the regression coefficient in the linear regression, and to the feature importance value in the Random Forest regression. For each predictor and FCR we plot the maximum importance value achieved from the 12 lags considered (except for the emissions that only consider the 1 month lag). Overall, both methods yield similar regions with high/low predictive power (i.e. importance values) for all the predictors. Lagged emissions show by large the highest predictive power worldwide, with the AMO showing by contrast the lowest predictive role, especially for the Random Forest regressions. From the rest of the CIs, the Indian ocean indices, the PDO and the NAO show many regions worldwide with relatively large importance values, which

extend beyond their typically reported areas of influence. By contrast, El Niño indices (and especially those defined more to the east) tend to show comparatively weaker importance values all around the world, including in the American continent, where it is known to exert an important influence. This might be due to the strong seasonality of ENSO influences, which cannot be fully captured with our methodological approach. It is also possible that, given the short time span of the study period, the indices exhibiting larger weights are those which happen to have similar trends than the emissions data, which would explain why indices like ENSO, that mostly varies at shorter (interannual) timescales, have comparatively weaker importance.

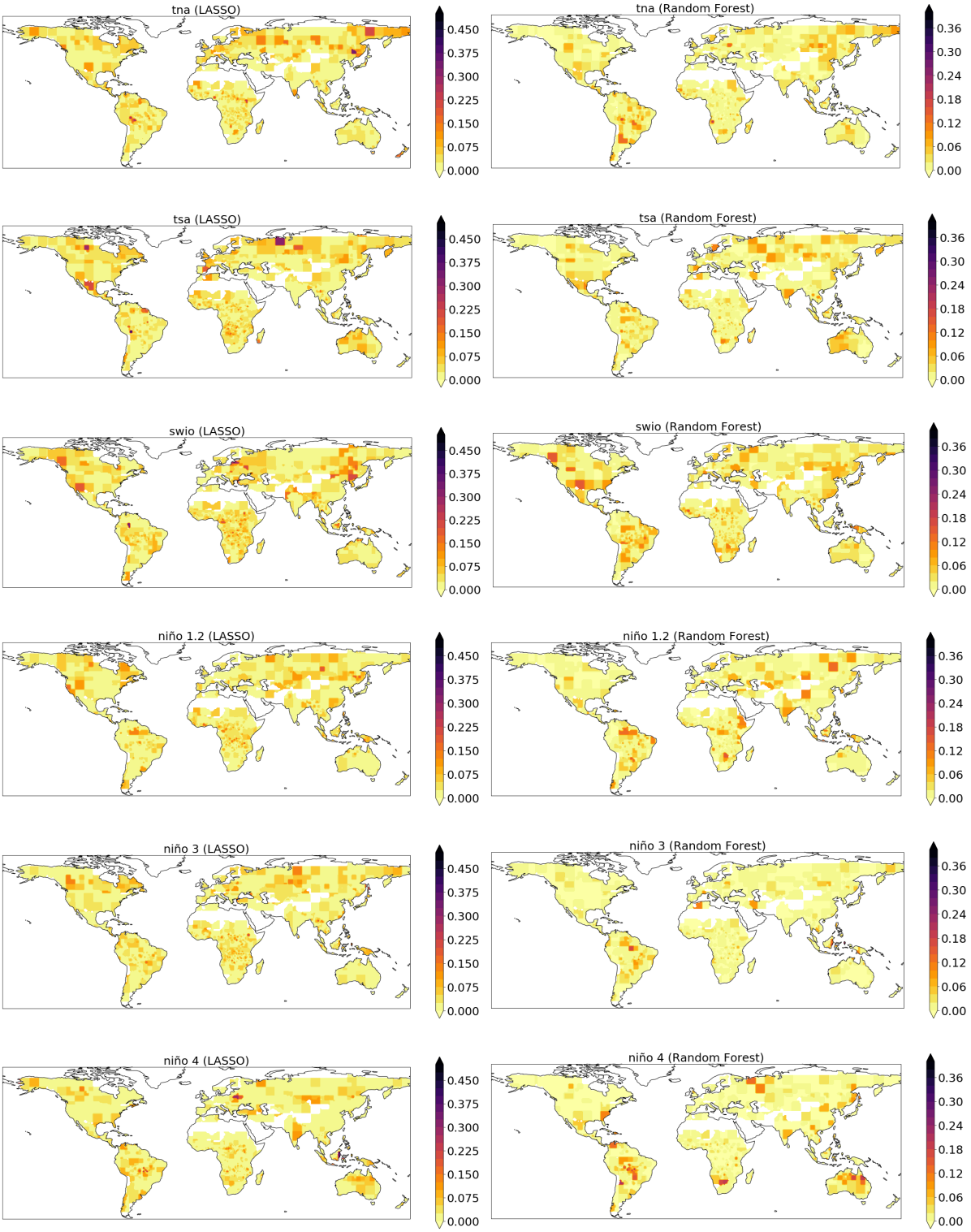

Figure 23: Predictor importance, represented as the absolute value of the regression coefficients (for the linear regression with LASSO regularisation; left) or the feature importance (for the optimised Random Forest regression; right) for each predictor, considering the lag for which it is maximum. All values are derived for the regressions using CIs (lagged by 1 to 12 months) and 1-month lagged emissions as predictors.

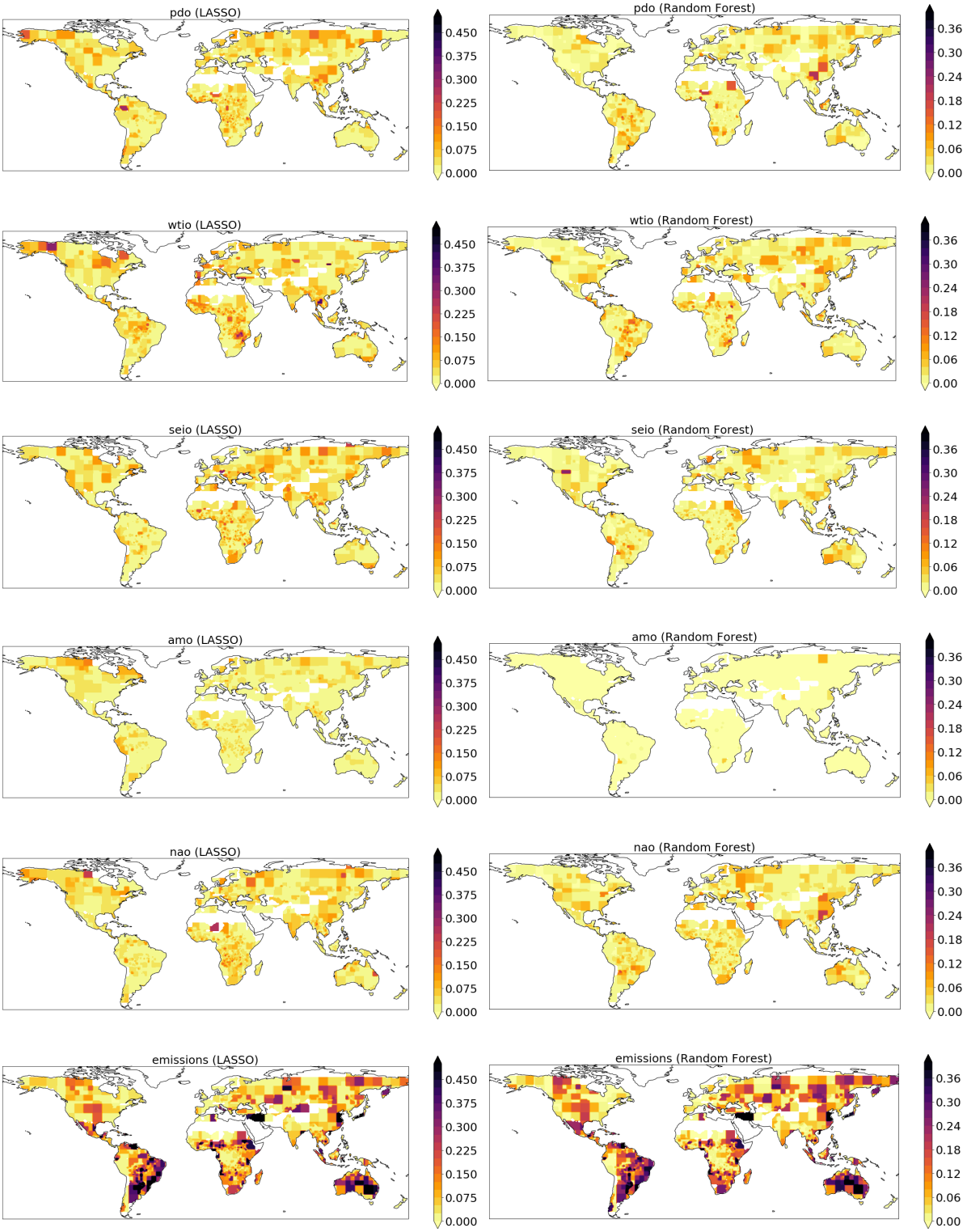

Figure 23: (Continuation)

5 Conclusions

5.1 Main results and lessons learned

- Due to the large availability of predictors and the relatively short time span of the observational datasets, the use of simple linear regressions led to strong overfitting. Introducing a method that identified and penalised the spurious predictors, like the LASSO regularisation technique, was needed to overcome this problem. For the non-linear approaches (i.e. the Random Forest regression), a key methodological decision to minimise overfitting and improve the actual performance was to lower to 2 the maximum permitted depth for each of the decision trees.
- The main result of the study is that the addition of lagged emissions greatly improved the performance of the regression models, especially in areas where CIs showed no predictive capacity. Only when both lagged emissions and CIs were considered did the model beat the climatological benchmark in a majority of FCRs.
- The predictive capacity of the methods was found to vary considerably between FCRs. Some FCRs (e.g. Australia, the Amazon basin) generally showed good performance scores while other FCRs showed negative or rather weak positive test scores (e.g. Northern North America). A compromise solution to minimise the chances of performing spurious predictions could be to use climatological emissions in FCRs in which the regression models consistently show poor testing scores (that is, only using the regressions on the regions that work effectively).
- Land cover features did not show any useful predictive relationships with the biomass burning emissions.
- The linear model with a LASSO regularisation and 1-month lagged emissions added as a predictor was the best performing model, improving the climatology benchmark on more than half of the total emissions and outperforming climatology in general.
- To introduce some seasonality in the predictions, we did some tests including the months as predictors. However, for this rather simplistic approach the coefficients associated with these seasonality predictors were completely negligible in both the regularised linear model and the Random Forest regression model. The appropriate way to take seasonality into account would be to build different regression models for the biomass burning emissions of each month of the year, something that we tried but that yielded very poor performance because it reduced the training sample to 15 points (one for each calendar year not included as holdout). Such models could be built in the future when much longer datasets of emissions data become available.

5.2 Ideas for the follow-up

- We could aggregate the emissions over larger spatial domains to reduce the noise, and thus retain more predictable signals. Indeed, some of the Fire Cohesive Regions currently considered can be too small compared with the areas that are typically affected by CI variations via atmospheric teleconnections. Geographical regions defined in GFED (Global Fire Emissions Database) were considered, but those regions are, on the contrary, too big

and not self-consistent like the FCRs. Defining regions of an intermediate size that still describe self-consistent emissions (e.g. by merging nearby regions whose emissions correlate significantly) would be the ideal way forward.

- The ad-hoc score considered in this analysis to evaluate the quality of the regression models is not as easily interpretable as other more common metrics, such as temporal correlations or the mean square error skill score, in particular when comparing results from different methodologies. To be able to apply more standard metrics, we are considering the implementation of leave-one-out cross-validation, which also has the advantage of providing a single test score that accounts for the model performance for the whole study period.
- Introducing standard metrics would also help to assess the statistical significance of the results, and identify the FCRs for which the regression models are truly skilful.
- A detailed analysis of the stability of the regression coefficients/weights across the different holdouts could provide additional information on the regions and regression methods that yield more reliable results.
- Predictions with 4-month-lagged emissions and CIs lagged from 1 to 12 months (which we performed but did not show) had better performance that those in which the CIs lags from 1 to 3 months are not considered. If this improved performance is confirmed to be statistically significant, future developments could consider the use of CI predictions from the operational seasonal forecasts to provide the information of the months corresponding to lags 1 to 3 when making predictions of biomass burning emissions 4 months ahead, provided that the seasonal predictions are skilful for those CI. This approach, however, would need additional work to determine how to take into account the uncertainty in the CI values in the ensemble predictions, and whether the forecast CIs would require a recalibration of the regression models.

6 References

Breiman, L. (2001). Random Forests. *Machine Learning.* 45, 5–32 https://doi.org/10.1023/A:1010933404324

Chen, Y., Morton, D. C., Andela, N., Giglio, L., & Randerson, J. T. (2016). How much global burned area can be forecast on seasonal time scales using sea surface temperatures? *Environmental Research Letters*, 11(4), 045001. https://doi.org/10.1088/1748-9326/11/4/045001

Chen, Y., Randerson, J. T., Coffield, S. R., Foufoula-Georgiou, E., Smyth, P., Graff, C. A., et al. (2020). Forecasting global fire emissions on subseasonal to seasonal (S2S) time scales. *Journal of Advances in Modeling Earth Systems*, 12, e2019MS001955. https://doi.org/10.1029/2019MS001955

Chen, Y., Randerson, J. T., Morton, D. C., DeFries, R. S., Collatz, G. J., Kasibhatla, P. S., Giglio, L., Jin, Y., & Marlier, M. E. (2011). Forecasting fire season severity in South America using sea surface temperature anomalies. *Science*, 334(6057), 787–791. https://doi.org/10.1126/science.1209472

Coscarelli, R., Aguilar, E., Petrucci, O., Vicente-Serrano, S.M., Zimbo, F. (2021). The Potential Role of Climate Indices to Explain Floods, Mass-Movement Events and Wildfires in Southern Italy. *Climate*, 9, 156. https://doi.org/10.3390/cli9110156

Fernandes, K., Baethgen, W., Bernardes, S., DeFries, R., DeWitt, D. G., Goddard, L., Lavado, W., Lee, D. E., Padoch, C., Pinedo-Vasquez, M., & Uriarte, M. (2011). North tropical Atlantic influence on western Amazon fire season variability. *Geophysical Research Letters*, 38, L12701. https://doi.org/10.1029/2011GL047392

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), 1.

Pedregosa, F. (2011). Scikit-learn: Machine Learning in Python. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

Uyanık, T., Karatuğ, Ç., Arslanoğlu, Y. (2020). Machine learning approach to ship fuel consumption: A case of container vessel. *Transportation Research Part D: Transport and Environment*, Volume 84, 102389. https://doi.org/10.1016/j.trd.2020.102389

Zhou, S., Mentch, L. (2022). Trees, forests, chickens, and eggs: when and why to prune trees in a random forest. *Stat. Anal. Data Min.: ASA Data Sci. J.*. , 1–. https://doi.org/10.1002/sam.11594

Document History

Version	Author(s)	Date	Changes
Version 1.0	BSC	14/10/2022	First draft sent for review
Version 1.1	BSC	27/10/2022	Final version addressing the comments from the reviewers

Internal Review History

Internal Reviewers	Date	Comments
CNR-ISAC	24/10/2022	
MéteoFrance	25/10/2022	

Estimated Effort Contribution per Partner

Partner	Effort
BSC	7 PMs
Total	**7 PMs**