Data Descriptor: Estimating nitrogen and phosphorus concentrations in streams and rivers across the contiguous United States: a machine learning framework

Longzhu Shen[1+] & Giuseppe Amatulli[2,3+*], Tushar Sethi [4], Peter Raymond [2], Sami Domisch [5]

March 13, 2019

1. University of Cambridge, Department of Zoology, Cambridge, CB2 3EJ, UK
2. Yale University, School of Forestry & Environmental Studies, New Haven, CT 06511, USA.
3. Yale University, Center for Research Computing, New Haven, CT 06511, USA.
4. Spatial-Ecology, Meaderville House, Wheal Buller, Redruth TR16 6ST, UK.
5. Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Department of Ecosystem Research, 12587 Berlin, Germany.

[*]Corresponding author: Giuseppe Amatulli (giuseppe.amatulli@gmail.com)
[+]Equal contribution

Abstract

Nitrogen (N) and Phosphorus (P) are essential nutrients for life processes in water bodies but in excessive quantities, they are a significant source of aquatic pollution. Eutrophication has now become widespread due to such an imbalance, and is largely attributed to anthropogenic activity. In view of this phenomenon, we present a new dataset and statistical method for estimating and mapping elemental and compound concentrations of N and P at a resolution of 30 arc-seconds ($\sim$1 km) for the conterminous US. The model is based on a Random Forest (RF) machine learning algorithm that was fitted with environmental variables and seasonal N and P concentration observations from 230,000 stations spanning across US stream networks. Accounting for spatial and temporal variability offers improved accuracy in the analysis of N and P cycles. The algorithm has been validated with an internal and external validation procedure that is able to explain 70-83% of the variance in the model. The dataset is ready for use as input in a variety of environmental models and analyses, and the methodological framework can be applied to large-scale studies on N and P pollution, which include water quality, species distribution and water ecology research worldwide.

Background & Summary

Nitrogen (N) and phosphorus (P) are primary nutrients and vital for life processes such as protein synthesis, cellular growth and reproduction. However, in inordinate quantities, the two elements are also a major source of stream and river impairment [1]. Large inputs of these limiting nutrients can cause deleterious algal growth with a myriad of negative ecosystem responses including eutrophication [23, 50]. In rivers, the pre-anthropogenic concentrations and fluxes of N and P were generally small and much less than the present day, with inputs stemming mainly from erosion and the leakage of dissolved organic N and P [11, 49]. However, the increased presence of these two elemental pollutants in many rivers is now owed to anthropogenic activity such as fertiliser use, increased output from waste-water treatment plants and atmospheric nitrogen deposition [9, 41]. This has led to the widespread eutrophication of both inland and coastal waters [10].

Over the past decades, significant progress has been made towards our understanding of the dynamics of natural and anthropogenic inputs of N and P to inland waters. Furthermore, the recognition of human impact on the N and P cycle has driven much research into the scope for better management of these nutrients [8, 10].

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27585v1 | CC BY 4.0 Open Access | rec: 13 Mar 2019, publ: 13 Mar 2019

However, our current ability to map N and P concentrations across regions or the globe is still limited. Early attempts focused on concentrations and fluxes from major rivers [9, 34] but were implemented through bottom up approaches, which estimated N and P content based on our knowledge of land-use and population centre influences on river nutrients [19, 30, 47, 31]. Other local and regional studies have also featured different combinations of bottom up, process based, and statistical models, which link N concentrations in inland water to environmental variables. [16, 24, 46, 53].

Freshwater environmental variables (climate, topography, land cover, surface geology and soil) that account for the basin and upstream environment have recently been computed [12]. This set of stream variables at the near-global scale provides a new base for stream-relevant biotic and abiotic modelling, such as biodiversity composition, nutrient distribution, or water flow. Based on this platform, we present in this paper a new method of mapping the concentration of N and P across continental waters, which employs a statistical approach within a machine learning framework. The resulting N and P maps can be used in studies focusing on nutrient loading and processing in inland waters. For instance, fertiliser run-off presents a big loss of chemical nutrients to recipient fresh water bodies, and can be charted by the aforementioned method [18, 45]. The N and P maps possess information about the location of nutrient-enriched streams, which can guide engineered de-nitrification processes [43, 48]. In addition to resource recovery, a mitigation strategy can be employed through the improved management of nutrient-rich wastes. In this approach as well, the derived N/P ratio map can prove a valuable source of information. Furthermore, this unique N and P modelling can be used in conjunction with process-based methods to enhance the understanding of riverine N and P processes.

In this paper, we present a dataset derived by connecting freshwater environmental variables with *in situ* measurements for mapping the distribution of N and P, and their various compounds, in water bodies across the conterminous US. The statistical model is based on Random Forest (RF)[7], a well-established machine learning approach with the capability of handling complex and heterogeneous data. We demonstrate in detail below how RF has excelled to date at capturing local geographical variations of stream predictors, and produced superior predictive performance for N and P distribution in the US. The mapped resolution of the predicted N and P concentrations is at a 30 arc-second gridded stream network (∼1 km) for four seasons. The described dataset (Data Citation 1) is ready for use as input data in various environmental models and analyses. The newly developed dataset and the methodological framework are suitable for large-scale environmental analyses such as N and P emissions in small and large rivers at a global scale. To our knowledge, this is the first time that N and P concentrations have been estimated at such high spatial resolution, while accounting for spatial and temporal variability, which reflect N and P cycles more accurately.

Methods

In this study, we constructed models using two data sources: measured concentrations for N and P (referred hereafter as observations or response variables) provided by the United States Geological Survey (USGS), and stream variables (referred hereafter as predictors) computed by Domisch et. al 2015. In the following section, we describe in sequence the data source and the modelling framework.

Source dataset

N and P concentration observations

We retrieved from the USGS water portal (`https://www.waterqualitydata.us/portal`) the measured concentration data for N and P nutrients in their various forms for the period 1994-2014. The data was collected by more than 230,000 stations spanning US stream networks. Each single observation is associated with its sampling geolocation (latitude and longitude) and a USGS Parameter Code (PC) for its chemical identity. We selected five nutrients of interest as the response variables. The PC descriptions of the chemical nature of each nutrient, as well as abbreviations are summarised in Table 1.

Table 1: Chemical nutrients with their USGS Parameter Code (PC) and abbreviation.

PC	Description	Abbreviation
00600	Total Nitrogen	TN
00665	Total Phosphorus	TP
00602	Total Dissolved Nitrogen	TDN
00666	Total Dissolved Phosphorus	TDP
00618	Nitrate	NO3

We prepared a monthly average of the concentration of each nutrient for each sampling station. Then, we aggregated these monthly values to a seasonal level using the median of three consecutive months corresponding to a season, as defined in Table 2. The final aggregated results, i.e. the number of observations, are sufficiently large (see Table 2) to build machine learning based models.

Table 2: Number of observations of the nutrients for each of the four seasons.

Season	1 (Winter)	2 (Spring)	3 (Summer)	4 (Autumn)
Month	11-12-01	02-03-04	05-06-07	08-09-10
TN	3252	4506	5156	5363
TDN	2125	2775	3905	3619
NO3	4347	5817	9810	10140
TP	4534	5768	7480	7793
TDP	3074	3892	6050	5902

Stream predictor layers

To build the predictive models, we used a total of 47 predictors belonging to four categories: topography, soil, land cover and climate (Table 3). All predictors are considered freshwater-specific environmental variables [12] that account for the upstream characteristics of the watershed and longitudinal connectivity across the 30 arc-second HydroSHEDS stream network [25]. For each 30 arc-second grid-cell along the stream network, the upstream catchment and stream was delineated, i.e., where each grid-cell serves as a virtual pour-point, and then overlaid with range-wide environmental layers (Table 3). This yielded a series of predictors such as the upstream average forest cover, upstream sum of precipitation that mimics surface run-off and the average upstream temperature [12].

All predictors except climate were considered static, as opposed to being time-updated. Monthly climate data was aggregated (mean aggregation) to a seasonal level, where winter was specified as the time frame December-February, spring as March-May, summer as June-August, and autumn as September-November. Regarding temperature layers, we only aggregated the upstream air temperature across the upstream cells, while all other layers were aggregated across the entire sub-catchment [12]. The unit for each stream variable is derived from an original, spatially continuous environmental variable across the land area. Thus, temperature is expressed in degrees Celsius, precipitation in millimetres, and land cover as a percentage of each class (e.g. Urban/built-up class in percentage). We refer to [12] for further details on the calculation of the freshwater-specific predictors.

The N and P observation points report latitude and longitude of the sampling location. Nonetheless, due to spatial inaccuracies in sampling locations of the HydroSHEDS stream network, the latitude and longitude locations do not consistently fall directly on the streams. We therefore employed a "snapping" procedure to move all observation points to their closest stream cell using the *r.stream.snap* function in GRASS GIS [14] with 3 km as the maximum snapping distance.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27585v1 | CC BY 4.0 Open Access | rec: 13 Mar 2019, publ: 13 Mar 2019

Table 3: Overview of all 47 environmental predictors used in the models. Please see [12] for further information on each predictor.

Variable type	Variable name	Variable description	Citation
elevation	dem	Average elevation	[25]
slope	slope	Average slope	
topology	ord	Stream order	
soil	soil01	Soil organic carbon	[20]
	soil02	Soil pH in H2O	
	soil03	Sand content mass fraction	
	soil04	Silt content mass fraction	
	soil05	Clay content mass fraction	
	soil06	Coarse fragments (> 2 mm fraction) volumetric	
	soil07	Cation exchange capacity	
	soil08	Bulk density of the fine earth fraction	
	soil09	Depth to bedrock (R horizon) up to maximum 240 cm	
	soil10	Probability of occurrence (0-100%) of R horizon	
land cover	lc01	Evergreen/deciduous needleleaf trees	[51]
	lc02	Evergreen broadleaf trees	
	lc03	Deciduous broadleaf trees	
	lc04	Mixed/other trees	
	lc05	Shrubs	
	lc06	Herbaceous vegetation	
	lc07	Cultivated and managed vegetation	
	lc08	Regularly flooded shrub/herbaceous vegetation	
	lc09	Urban/built-up	
	lc10	Snow/ice	
	lc11	Barren lands/sparse vegetation	
	lc12	Open water	
temperature	tmin	Monthly temperature average min	[22]
temperature	tmax	Monthly temperature average max	
precipitation	prec	Sum of monthly precipitation	
hydrology and climate	hydro01	Annual Mean Upstream Temperature	
	hydro02	Mean Upstream Diurnal Range (Mean of monthly (max temp - min temp))	
	hydro03	Upstream Isothermality (hydro02 / hydro07) (* 100)	
	hydro04	Upstream Temperature Seasonality (standard deviation *100)	
	hydro05	Maximum Upstream Temperature of Warmest Month	
	hydro06	Minimum Upstream Temperature of Coldest Month	
	hydro07	Upstream Temperature Annual Range (hydro05 - hydro06)	
	hydro08	Mean Upstream Temperature of Wettest Quarter	
	hydro09	Mean Upstream Temperature of Driest Quarter	
	hydro10	Mean Upstream Temperature of Warmest Quarter	
	hydro11	Mean Upstream Temperature of Coldest Quarter	
	hydro12	Annual Upstream Precipitation	
	hydro13	Upstream Precipitation of Wettest Month	
	hydro14	Upstream Precipitation of Driest Month	
	hydro15	Upstream Precipitation Seasonality (Coefficient of Variation)	
	hydro16	Upstream Precipitation of Wettest Quarter	
	hydro17	Upstream Precipitation of Driest Quarter	
	hydro18	Upstream Precipitation of Warmest Quarter	
	hydro19	Upstream Precipitation of Coldest Quarter	

Modeling framework

Response variable transformation

The distribution of the observed nutrient (TN, TP, TDP, TDP, and NO3) data were all significantly skewed to the left. We calculated skewedness values by the *moments* R library using **skewness** command. The skewedness values were all larger than 8 (see Supplementary Figure A.1 and A.2). Such a level of skewedness may cause significant biases in the data training process. We therefore applied a Box–Cox power transformation [6] to make corrections. After transformation, data were more symmetrically distributed as indicated by a near zero skewedness and higher linear behaviour on the Q-Q plots (see Supplementary Figure A.1 and A.2). The Box–Cox power transformation maps a variable y_i to a transformed variable ${y_i}^{(\lambda)}$ defined by

$${y_i}^{(\lambda)} = \begin{cases} \frac{{y_i}^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0, \\ \log(y_i), & \text{if } \lambda = 0. \end{cases} \tag{1}$$

It is estimated that the transformed ${y_i}^{(\lambda)}$ will follow approximately normal distributions so that the data are more transferable to many modelling techniques. In our case, the transformed output did produce a more symmetric distribution although it did not reach perfect normality. One of the strengths of RF is its ability to deal with a non-normal data distribution, so the presence of slight non-normality does not preclude good modelling performance. Furthermore, to reduce the influence of extreme values, only data with the coefficient of variation (CV=standard deviation/mean)> 0.5 and within the range of (2.5, 97.5) percentile were retained for the downstream model building steps.

Model training

We employed the Random Forest (RF) regression algorithm implemented in the R package *randomForest* [26] to train the nutrient distribution models. RF regression is an ensemble learning strategy that elevates the collective predictive performance of a large group of weaker learners (regression trees). Two key elements contributing the superiority of the RF algorithm are bootstrapping aggregation (bagging) and random selection of variables. Bagging (bootstrap sampling from the train data) aims at reducing data noise through averaging. Data that is not included in the bag is called an out-of-bag (OOB) sample. Random drawing of variables improves variance reduction by reducing the intercorrelation between trees. OOB samples can be used to validate the model performance (equivalent to cross validation) and evaluate the variable importance. The variable importance is of great value in identifying the most influential variables that direct predictive outcomes and thus offer adaptive or intervention strategies in response to the modelled phenomena. One important feature of the RF algorithm is its relative resilience towards data noise due to the two mechanisms mentioned above. This technical advantage of RF directly benefits the analysis of environmental data. For the sake of independent evaluation of the model's predictive performance, the whole data set was split into two portions: 60% of the data for model training (training set) and 40% for model testing (testing set).

Model validation

The predicting performance on the training and testing sets provided complementary information for the model validation. Training primarily exhibits model robustness, i.e. stability and balance of model predictability in the presence of data shuffling. Testing measures the model performance on the unseen data during the training phase and addresses the concern of model overfitting. In this context we used the Pearson correlation coefficient (2) as the statistical metric to quantify the predictive performance of the models.

$$\rho_{X,Y} = \frac{XY - XY}{\sqrt{X^2 - X^2}\sqrt{Y^2 - Y^2}} \tag{2}$$

where X and Y are the predicted and observed vectors and $\cdot$ denotes the expectation value. We also plotted the observed vs. predicted values and fitted a linear model to estimate discordance from the 1:1 relationship.

After establishment of the predictive models, we investigated the contributions of each variable to the predicted outcomes by means of the "variable importance", an output from RF. The variable importance is presented as the percent increment of the Mean Square Error (MSE) of the predictions (measured by OOB estimations) resulting from variable permutations. Moreover, to depict the spatial relationship among the response variables and the most contributing predictors, we designed a so-called "local correlation" approach. Specifically, we generated local Pearson coefficients between a predictor and one of the nutrients (TN, TP, TDP, TDP, and NO3) in a moving circular window with a radius of 51 cells. These local correlation maps were coupled with nutrient concentrations as a bivariate tuple to illustrate the local correlation power and its relation to the quantity of the response variable. From these maps, one is able to read out the variation of predictors' influence in different geographical regions. To our knowledge, this is the first time that a "local correlation" has been performed to interpret the relationship between the RF output and its most important predictors.

Model prediction

We performed an inverted Box–Cox transformation on the predictive outcomes of the models to recover the physical values (expressed in ppm) for all the nutrients (TN, TP, TDP, TDP and NO3) within the conterminous US stream network for each of the 30-arc-second stream grid-cells.

Code availability

We used the following open source software packages to compute the N and P predictions:

- Geospatial Data Abstraction Library (GDAL, `http://www.gdal.org`, version number 2.1.2) [52].
- Geographic Resources Analysis Support System software (GRASS, `https://grass.osgeo.org`, version number 7.3.0) [36, 35].
- Processing Kernel for geospatial data (pktools, `http://pktools.nongnu.org`, version number 2.6.3)[32].
- R: a language and environment for statistical computing[39], with the following libraries: randomForest, quantregForest, geoR, plyr, moments, data.table, bit64.

All of these tools provide fast and scaleable computation features and functions for raster-based workflows that are easily automated using a scripting language, such as Bash or Python[2]. They also allow for the processing of very large datasets owing to efficient algorithms and optimised memory management. All calculations were processed in parallel using open-source software at the Centre for Research Computing of Yale University.
In the spirit of reproducible research we provide the scripting procedure at: `https://gitlab.com/Ferdinand18/nutr_us_streams`. The full procedure, starting from the N and P observations treatment to 30-arc-second raster predictions, consists of 5 nested codes:

- 01_Data_cleaning.sh: cleaning the raw observation data.
- 02_Snapping.sh: snapping the observation data points onto the river streams.
- 03_Extraction.sh: extracting descriptors corresponding to the snapped points.
- 04_US_Seasonal_Modelling.sh : building predictive models based on the observation data.
- 05_US_Prediction.sh: making predictions for all the US streams and building geotif maps as the final output.

Data Records

We provide TN, TDN, NO3, TP, and TDP concentrations (ppm) for four seasons (winter, spring, summer and autumn) for the gridded stream network at a spatial grain of 30 arc-second (∼1 km). All layers are available for download at Data Citation 1. The nutrient concentrations, mapped across the conterminous USA, are available in a compressed GeoTiff file format in the WGS84 coordinate reference system (EPSG:4326 code). All layers are stored as floating points (Float32 data type) to ensure sufficient precision for future use and analysis for varied purposes.

The predicted nutrient maps follow the layer name convention:

nutrient abbreviation_resolution_season.format

Below are two examples of the layer names for the two main nutrients product TN and TP

– TN_1KM_winter.tif: layer showing the Total Nitrogen at 30 arc-second spatial resolution for the winter season.

– TP_1KM_summer.tif: layer showing the Total Phosphorus at 30 arc-second spatial resolution for the summer season.

For the purpose of visual interpretation of the results, we plotted the TN and TP bivariate maps as shown in Figure 1. The bivariate TN-TP map representation permits an immediate perception of the spatial patterns of these two nutrients in the same map. This visual result was achieved by a mean-value aggregation of the original 30 arc-second resolution nutrient distributions using a moving window of 10x10 grid-cells so that a continuous surface could be easily mapped across the entire conterminous US. Figure 1 shows high concentrations of TN and TP (red colour) in intensive agriculture/grazing areas (e.g. of the US Midwest) and also close to large urban areas (e.g. New York, Philadelphia, Baltimore, Washington DC). On the other hand, low concentrations of TN and TP are located in forestry/mountain areas (e.g. Rocky Mountains, Appalachian Mountains). This observation is in line with the anthropogenic eutrophication effect that coincides with intensive grazing and agricultural activities [17].

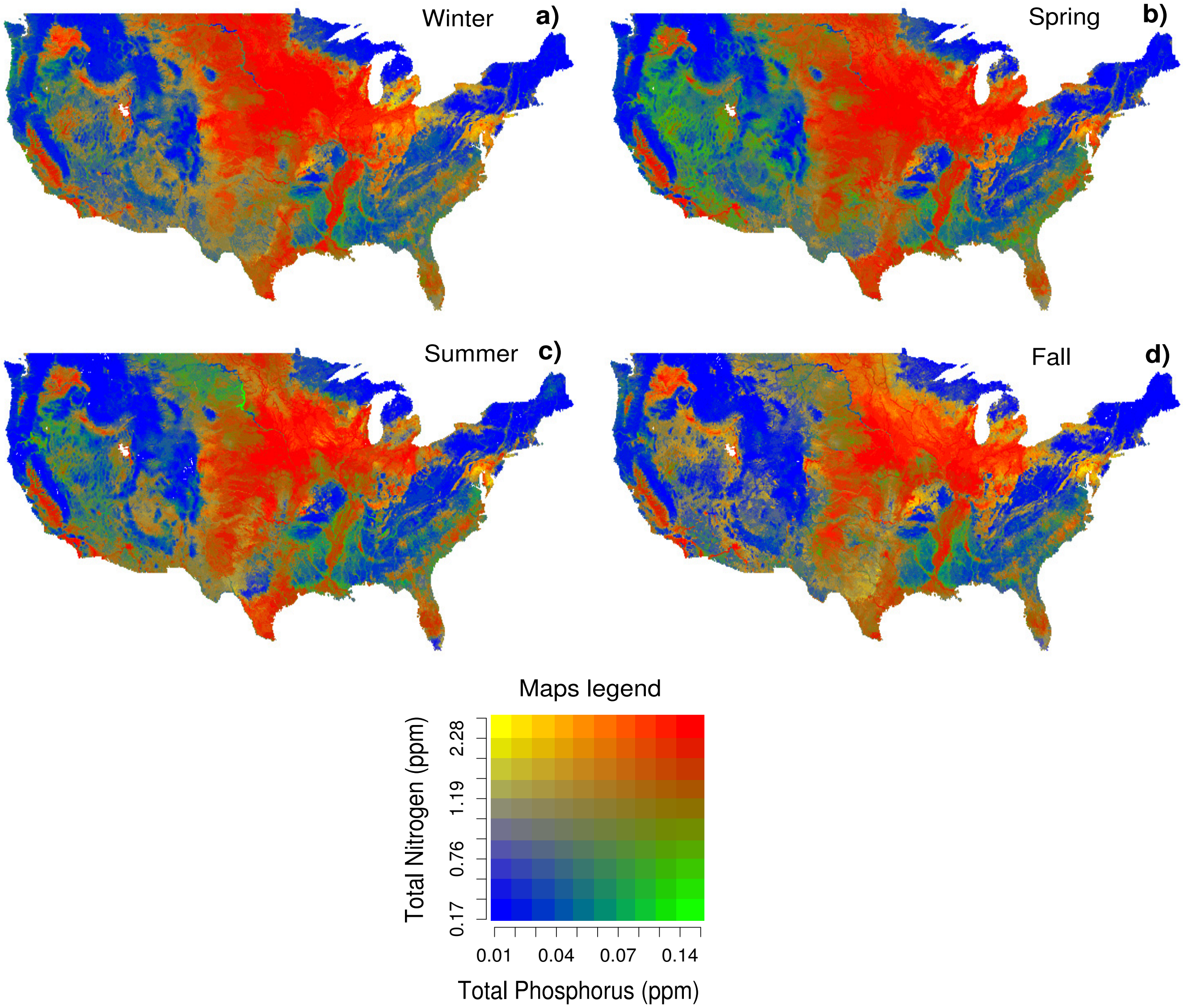


Figure 1: Bivariate maps showing the predicted Total Nitrogen (TN) and Total Phosphorus (TP) values across the four seasons. Streams and rivers on the original 30 arc-second resolution maps were aggregated using the mean value of a moving window with 10x10 grid-cells for a better visual effect. Red indicates high concentration areas, which mainly coincide with high agriculture or grazing activities or urban zones. Blue indicates low nutrient load areas, which are frequently occupied by forests or deserts.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27585v1 | CC BY 4.0 Open Access | rec: 13 Mar 2019, publ: 13 Mar 2019

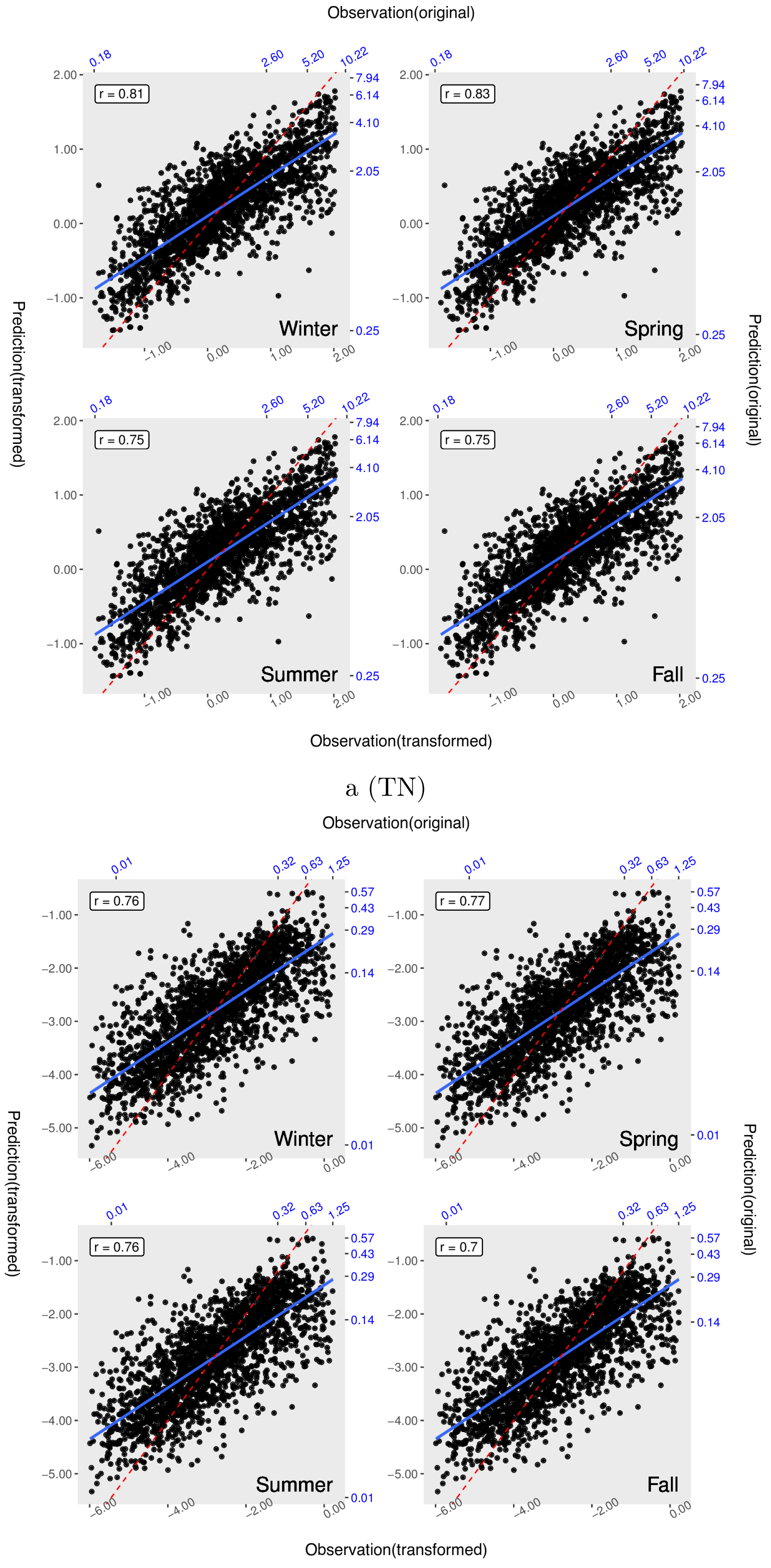
Technical Validation

The Pearson correlations between predicted and observed values for TN and TP are in the range of 0.7-0.83 across the testing sets as shown in Figure 2. The red dotted lines represent the the 1:1 relationship for each panel. The solid blues lines showed the regression of the black data points (prediction vs observation). The deviation between these two lines is sourced from the noise of the observation data and the predictor bias and variance [15, 13, 27]. Similar plots were generated for TDN, TDP and NO3 (see Supplementary Figure A.6). The high-level correlation for each plot and overall consistency among all of them suggest the appropriateness of the model fit. Moreover, the Pearson correlations for the training and testing are comparable, and free of model overfitting. The correlation graphs for the training set (TN, TP,TDN, TDP and NO3) are provided in the supplementary material (Figure A.7 and A.8).

After the establishment of the predictive models, we investigated the contributions of each variable to the predicted outcomes, by means of the "variable importance", which is an output of RF. As shown in Figure A.9 and A.10, we noticed two out of the three variables "urban/built-up", "cultivated and managed vegetation" (mainly agriculture fields) and "soil pH in H2O" were frequently ranked as the two most important contributors. This interesting observation drove our quantitative exploration of the impact of these variables using the "local correlation" approach. The local Pearson correlation map was coupled with the nutrient concentration by means of bivariate maps, as illustrated in the panels a) to d) in Figure 4 and 5.

From these maps, one is able to view the geographical regions that are influenced by the variable of interest in varying degrees. For instance the red zones on the maps are where the variable of interest has heavy positive impact and the nutrient concentrations are also high. In addition to the bivariate plots, we also counted the number of grid-cells that showed significant positive and negative correlations ($|r|>0.6$) for a particular variable and plotted them for all four seasons in the panel e) of Figure 4 and 5. From these plots, a seasonable pattern for the highly positive correlations ($r > 0.6$) becomes obvious. Specifically, the variable of interest has more weights on its correlation with the observed concentration of a nutrient during spring and summer than those on autumn and winter (right sub-panel of panel e). This phenomenon can be conceptually comprehended by the fact that higher biological and anthropological activities during spring and summer than those are in the autumn and winter.

Figure 2: Scatter plots regarding the model performance (testing dataset) for a) TN and a) TP, where each figure in the panel corresponds to a season. Scatter plots of TDN, TDP and NO3 are reported in Figure SI A.7 A.6



a (TN)

b (TP)

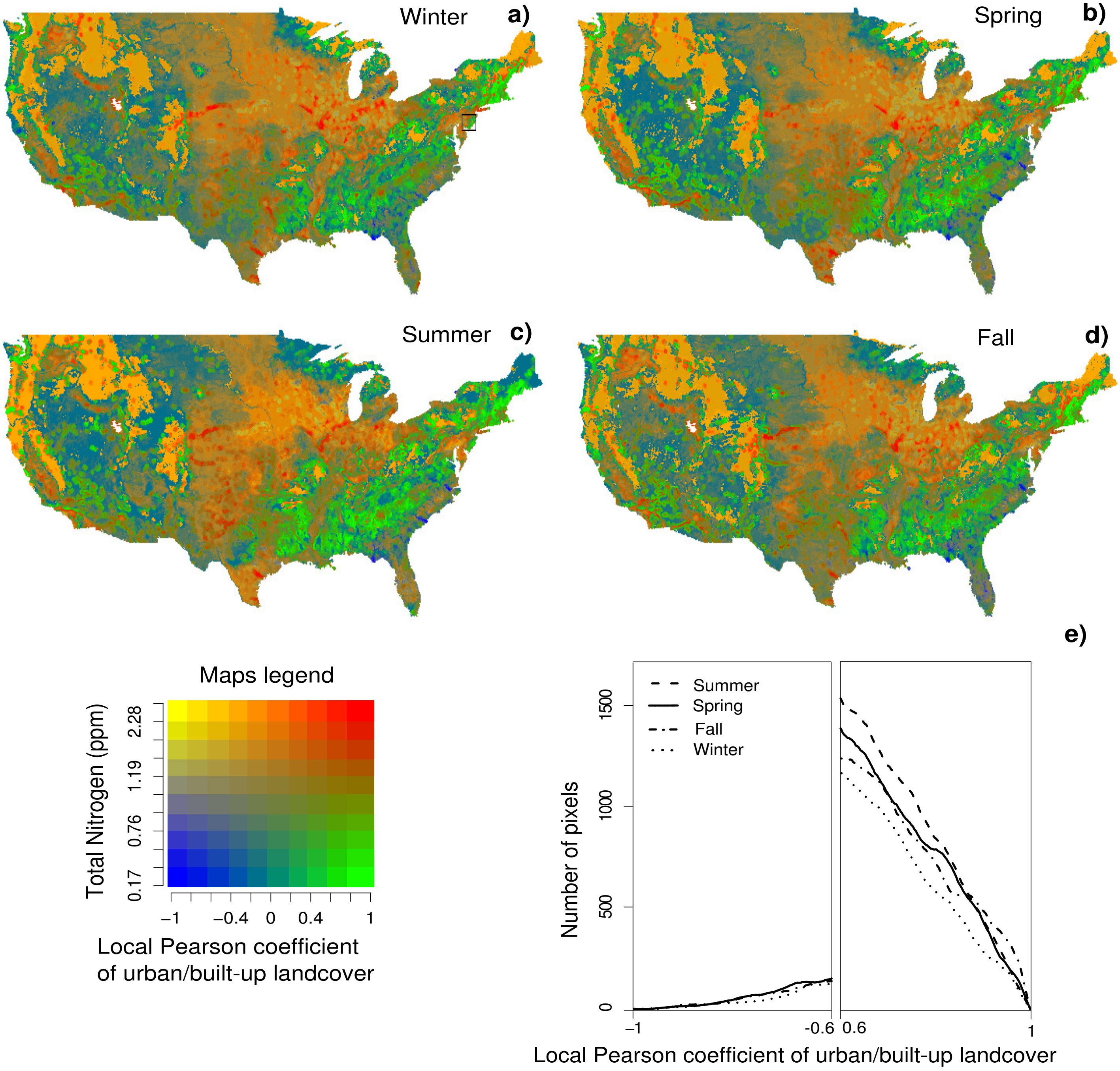


Figure 3: a-d) Bivariate maps representing the relationship between the Total Nitrogen (TN) and the local Pearson coefficient of (1st important variable) urban/built-up landcover vs TN. e) Graph showing the histogram tails of the raster map of the local Pearson coefficient. The maps show the influence/correlation of the urban/built-up (expressed in percentage) to the level of TN. A positive high correlation may appear in areas with low levels of TN (green), or areas with high levels of TN (red). Similarly, negative correlations could co-occur in low concentration ares (blue) or in high concentration areas (yellow). Streams and rivers are aggregated by a moving window of 10x10 grid-cells depicting the mean value for a better visualisation effect. The e) graph highlights a stronger positive correlation of the urban/built-up during warmer seasons (summer and spring) comparison with the colder seasons (autumn and winter). The black rectangle in the a) plot is zoomed in Figure A.12

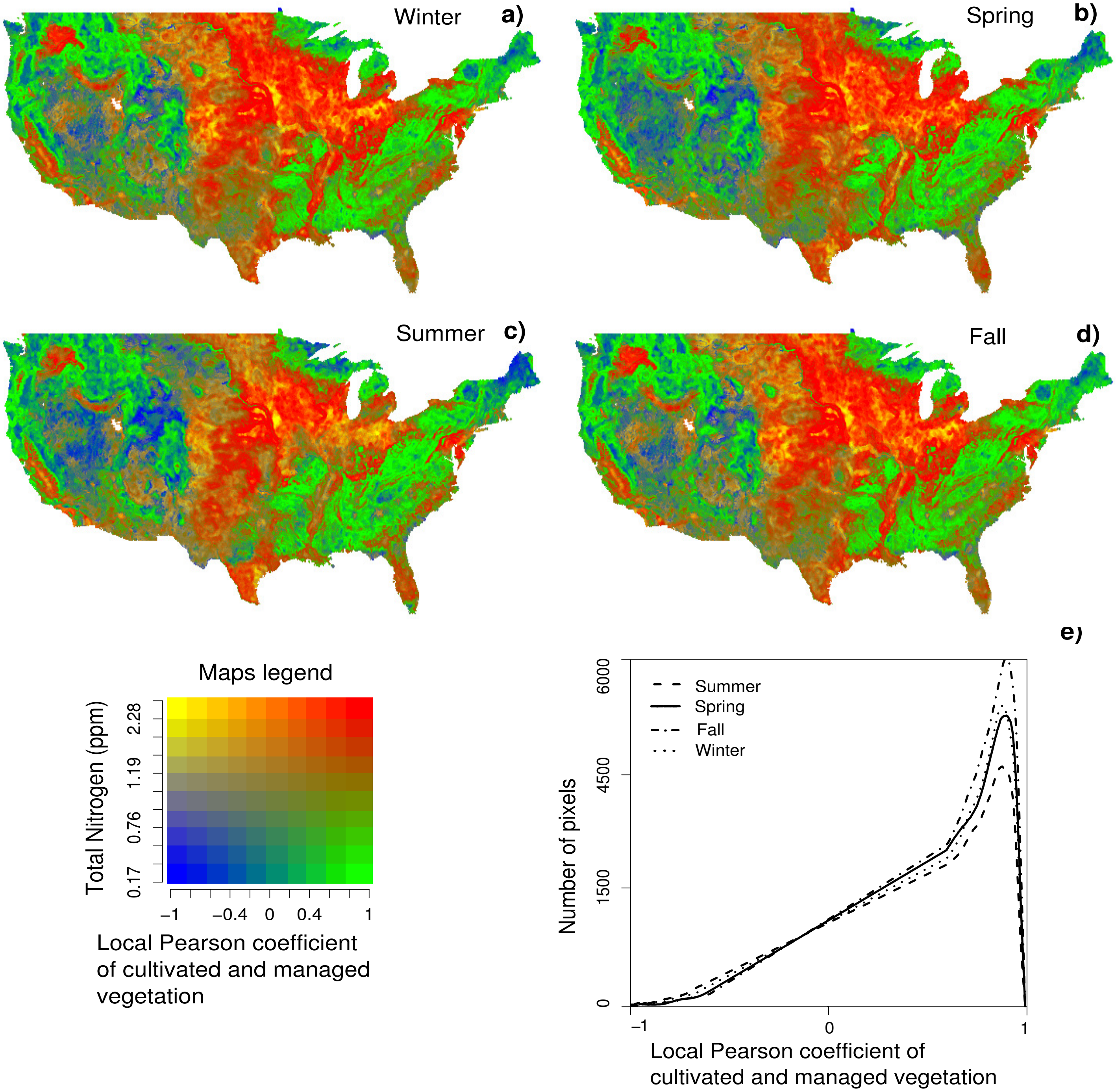


Figure 4: a-d) Bivariate maps representing the relationship between Total Nitrogen (TN) and the local Pearson correlation coefficient of cultivated and managed vegetation (2nd most important variable) vs TN. e) Graph showing the entire histogram of the raster map of the local Pearson coefficient. The maps show the influence/correlation of the cultivated and managed vegetation (expressed in percentage) to the level of TN. A positive high correlation could appear in areas with low levels of TN (green), or in areas with high levels of TN (red). Similarly, negative correlations could co-occur in low concentration areas (blue) or in high concentration areas (yellow). Streams and rivers are aggregated by a moving window of 10x10 grid-cells depicting the the mean value for a better visualisation effect. The graph highlights a stronger positive correlation of the cultivated and managed vegetation through all seasons, especially winter and autumn.

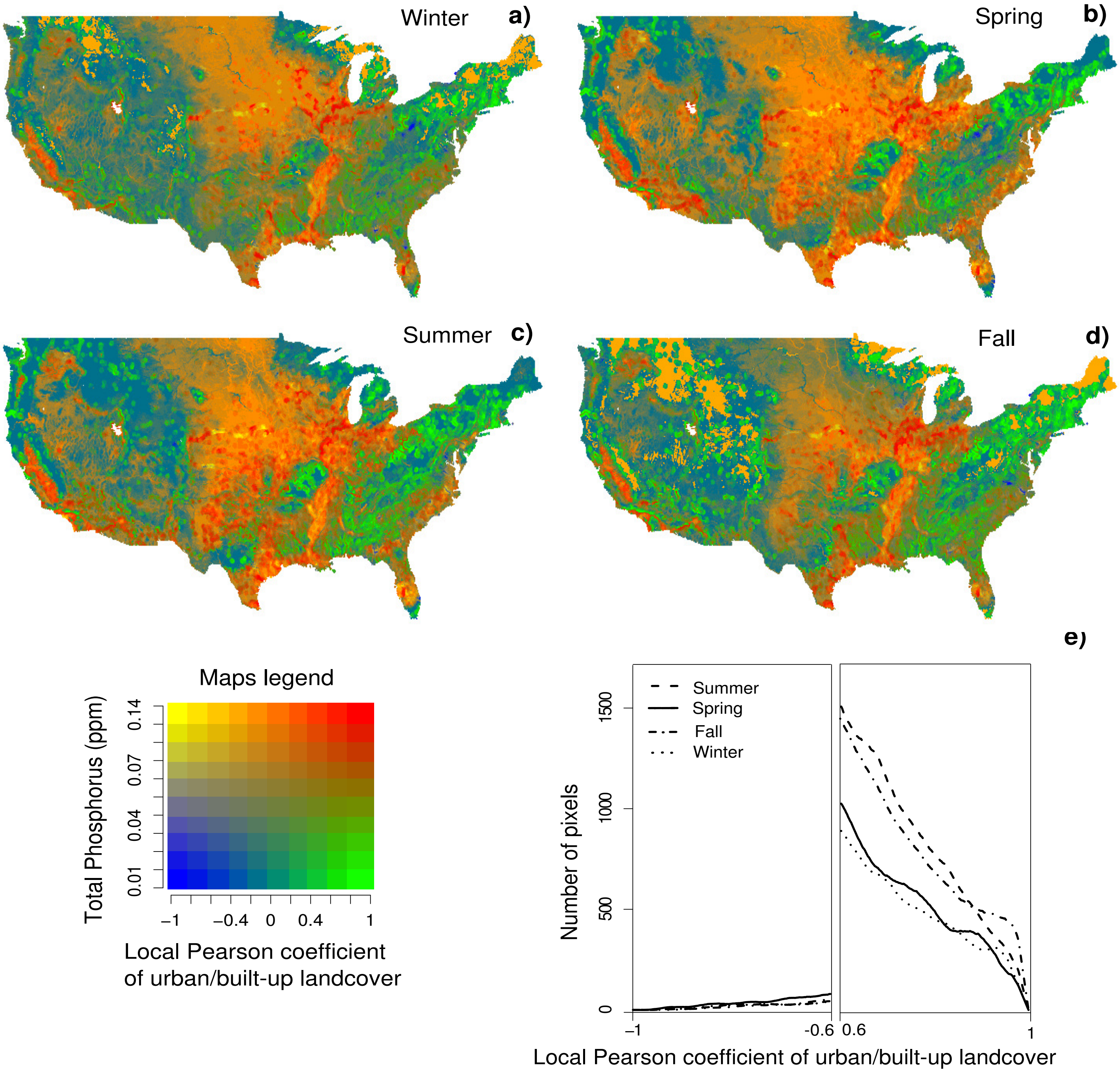


Figure 5: a-d) Bivariate maps representing the relationship between the Total Phosphorus (TP) and the local Pearson coefficient of urban/built-up landcover (1nd important variable) vs TP. e) Graph showing the histogram tails of the raster map of the local Pearson correlation coefficient. The maps show the correlation of the Urban/built-up layer (expressed as a percentage) to the level of TP. A positive high correlation could appear in areas with low levels of TP (green) or in areas with high levels of TP (red). Similarly, negative correlation could co-occur in low-concentration areas (blue) or in high-concentration areas (yellow). Streams and rivers are aggregated by a moving window of 10x10 grid-cells depicting the mean value for better visualisation. The graph highlights a stronger positive correlation of the urban/built-up for warmer seasons (autumn and summer), in comparison with colder seasons (winter and spring).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27585v1 | CC BY 4.0 Open Access | rec: 13 Mar 2019, publ: 13 Mar 2019

Usage Notes

The newly-developed stream nutrient concentration layers (Data Citation 1) have a wide array of potential applications in stream ecology, biodiversity research, conservation science, and stream and lake restoration ecology. For instance, the layers can be used to quantify the overall mass of of N and P discharged into specific lake or ocean bodies, enabling a deeper understanding of global-scale eutrophication[40]. Furthermore, these statistical estimates of nutrient concentration can be used to verify new process-based models that predict nutrient concentrations and transformations in inland waters worldwide [29]. The estimates can also be combined with maps of soil nutrient levels and fertiliser use to obtain information on terrestrial-aquatic coupling[44, 28]. Finally, the stoichiometry of the N/P ratio in natural/ecological systems is vital information for studying metabolic and biogeochemical processes. These new ratio maps can be used to enhance our knowledge on how coupled biogeochemical cycles impact ecosystems [38].

Overall, the newly-developed layers provide the basis for a variety of high-resolution, nutrient-related analyses across the inland waters in the conterminous US. A global-scale N and P assessment with new stream predictors at higher resolution (3-arc-second) is under development by the same group. The focus is on creating new geomorphometry variables (Geomorpho90m[5]) based on MERIT-DEM [54] by adopting the procedure described in[4]. The MERIT-DEM derived stream network is also under development[3]. These former described layers will be useful in combination with other global maps of irrigated areas [33], livestock[42], agricultural fertiliser use [37], soil types/properties[21] to compute N and P concentrations more accurately on a global scale. We encourage potential users of the described dataset to contact the authors for future product updates.

Acknowledgements

This study was supported in part by the facilities and staff at the Yale Centre for Research Computing (YCRC). Special thanks go to Dr. Peter Kempeneers, the developer of the pktools software, available at http://pktools.nongnu.org. His tools were fundamental in implementing a fast processing chain. G.A. also wishes to thank Dr. Anssi Pekkarinen who introduced him to the GeoComputation world under the Linux OS. S.D. was funded by the German Research Foundation DFG (grant DO 1880/1-1). This work benefited from discussions as part of the "Inland Water Global HydroBioGeoChemistry" working group. It was supported by NASA award NNX17AI74G and the Yale Institute for Biospheric Studies (YIBS).

Author Contributions

L.S., G.A., S.D. designed the study. L.S. and G.A. equally contributed to the manuscript by developing and implementing the computational methodology and the processing chain in the HPC cluster to estimate the N and P concentration, validated the dataset layers and wrote the first manuscript draft; S.D. provided important input on the processing chain; P.R. contributed to the observation data analysis; T.S. contributed in the manuscript drafting and final editing. All the authors contributed to the writing of the manuscript and interpretation of the results.

Competing financial interests

The author(s) declare no competing financial interests.

Data Citations

1. Shen, L. & Amatulli, G., et al. https://doi.pangaea.de/10.1594/PANGAEA.899168

References

[1] USEPA (US Environmental Protection Agency). National rivers and streams assessment 2008–2009: a collaborative survey, 2013.

[2] Giuseppe Amatulli, Stefano Casalegno, Remi D'Annunzio, Reija Haapanen, Pieter Kempeneers, Erik Lindquist, Anssi Pekkarinen, Adam M Wilson, and Raul Zurita-Milla. Teaching spatiotemporal analysis and efficient data processing in open source environment. In *Proceedingsofthe3rdOpen Source Geospatial-Research& Education Symposium*, page 13, 2014.

[3] Giuseppe Amatulli, Sami Domisch, Jens Kiesel, Tushar Sethi, Dai Yamazaki, and Peter Raymond. High-resolution stream network delineation using digital elevation models: assessing the spatial accuracy. Technical report, PeerJ Preprints, 2018.

[4] Giuseppe Amatulli, Sami Domisch, Mao-Ning Tuanmu, Benoit Parmentier, Ajay Ranipeta, Jeremy Malczyk, and Walter Jetz. A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. *Scientific data*, 5:180040, 2018.

[5] Giuseppe Amatulli, Daniel McInerney, Tushar Sethi, Peter Strobl, and Sami Domisch. Geomorpho90m - global high-resolution geomorphometry layers: empirical evaluation and accuracy assessment. *Scientific data*, 2019submited.

[6] George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252, 1964.

[7] Leo Breiman. Random Forests. *Mach. Learn.*, 45(1):5–32, October 2001.

[8] T. P. Burt, N. J. K. Howden, F. Worrall, M. J. Whelan, and M. Bieroza. Nitrate in united kingdom rivers: Policy and its outcomes since 1970. *Environ. Sci. Technol.*, 45(1):175–181, 2011.

[9] Nina F Caraco and Jonathan J Cole. Human Impact on Nitrate Export: An Analysis Using Major World Rivers. *Ambio*, 28(2):167–170, 1999.

[10] Walter K. Dodds, Wes W. Bouska, Jeffrey L. Eitzmann, Tyler J. Pilger, Kristen L. Pitts, Alyssa J. Riley, Joshua T. Schloesser, and Darren J. Thornbrugh. Eutrophication of U.S. Freshwaters: Analysis of Potential Economic Damages. *Environ. Sci. Technol*, 43(1):12–19, jan 2009.

[11] Walter K. Dodds, John R. Jones, and Eugene B. Welch. Suggested classification of stream trophic state: Distributions of temperate stream types by chlorophyll, total nitrogen, and phosphorus. *Water. Res*, 32(5):1455–1462, 1998.

[12] Sami Domisch, Giuseppe Amatulli, and Walter Jetz. Near-global freshwater-specific environmental variables for biodiversity analyses in 1 km resolution. *Scientific data*, 2, 2015.

[13] Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural. Comput.*, 4(1):1–58, 1992.

[14] GRASS Development Team. *Geographic Resources Analysis Support System (GRASS GIS) Software, Version 7.2.* Open Source Geospatial Foundation, 2017.

[15] Ulf Grenander. On empirical spectral analysis of stochastic processes. *Ark. Matemat.*, 1(6):503–531, Aug 1952.

[16] Bruna Grizzetti, Fayçal Bouraoui, and Alberto Aloe. Changes of nitrogen and phosphorus loads to European seas. *Glob. Chang. Biol*, 18(2):769–782, 2012.

[17] John A Harrison, Arthur HW Beusen, Gabriel Fink, Ting Tang, Maryna Strokal, Alexander F Bouwman, Geneviève S Metson, and Lauriane Vilmin. Modeling phosphorus in rivers at the global scale: recent successes, remaining challenges, and near-term opportunities. *Current Opinion in Environmental Sustainability*, 36:68–77, 2019.

[18] Murray R. Hart, Bert F. Quin, and M. Long Nguyen. Phosphorus Runoff from Agricultural Land and Direct Fertilizer Effects. *J. Environ. Qual.*, 33(6):1954, 2010.

[19] Bin He, Shinjiro Kanae, Taikan Oki, Yukiko Hirabayashi, Yosuke Yamashiki, and Kaoru Takara. Assessment of global nitrogen pollution in rivers using an integrated biogeochemical modeling framework. *Water Res.*, 45(8):2573–2586, 2011.

[20] T. Hengl, J.M. De Jesus, R.A. MacMillan, N.H. Batjes, G.B.M. Heuvelink, E. Ribeiro, A. Samuel-Rosa, B. Kempen, J.G.B. Leenaars, M.G. Walsh, and M.R. Gonzalez. Soilgrids1km - global soil information based on automated mapping. *PLoS ONE*, 9(8), 2014.

[21] Tomislav Hengl, Jorge Mendes de Jesus, Gerard BM Heuvelink, Maria Ruiperez Gonzalez, Milan Kilibarda, Aleksandar Blagotić, Wei Shangguan, Marvin N Wright, Xiaoyuan Geng, Bernhard Bauer-Marschallinger, et al. Soilgrids250m: Global gridded soil information based on machine learning. *PLoS one*, 12(2):e0169748, 2017.

[22] R.J. Hijmans, S.E. Cameron, J.L. Parra, P.G. Jones, and A. Jarvis. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25(15):1965–1978, 2005.

[23] Robert W. Howarth. Coastal nitrogen pollution: A review of sources and trends globally and regionally. *Harmful Algae*, 8(1):14–20, 2008.

[24] P. J. Johnes. Evaluation and management of the impact of land use change on the nitrogen and phosphorus load delivered to surface waters: The export coefficient modelling approach. *J. Hydrol.*, 183(3-4):323–349, 1996.

[25] B. Lehner, K. Verdin, and A. Jarvis. New global hydrography derived from spaceborne elevation data. *Eos*, 89(10):93–94, 2008.

[26] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[27] Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. *J. Am. Stat. Assoc.*, pages 101–474, 2002.

[28] Chaoqun Lu and Hanqin Tian. Global nitrogen and phosphorus fertilizer use for agriculture production in the past half century: shifted hot spots and nutrient imbalance. *Earth System Science Data*, 9(1):181–192, 2017.

[29] Taylor Maavara, Ronny Lauerwald, Goulven G Laruelle, Zahra Akbarzadeh, Nicholas J Bouskill, Philippe Van Cappellen, and Pierre Regnier. Nitrous oxide emissions from inland waters: Are ipcc estimates too high? *Global change biology*, 25(2):473–488, 2019.

[30] Emilio Mayorga, Sybil P. Seitzinger, John A. Harrison, Egon Dumont, Arthur H.W. Beusen, A. F. Bouwman, Balazs M. Fekete, Carolien Kroeze, and Gerard Van Drecht. Global Nutrient Export from WaterSheds 2 (NEWS 2): Model development and implementation. *Environmen. Model. Softw.*, 25(7):837–853, 2010.

[31] Michelle L McCrackin, Bärbel Muller-Karulis, Bo G Gustafsson, Robert W Howarth, Christoph Humborg, Annika Svanbäck, and Dennis P Swaney. A century of legacy phosphorus dynamics in a large drainage basin. *Global Biogeochemical Cycles*, 32(7):1107–1122, 2018.

[32] D. McInerney and P. Kempeneers. *Open Source Geospatial Tools - Applications in Earth Observation.* Springer Verlag, 2015.

[33] Jonas Meier, Florian Zabel, and Wolfram Mauser. A global approach to estimate irrigated areas–a comparison between different data and statistics. *Hydrology and Earth System Sciences*, 22(2):1119–1133, 2018.

[34] M. Meybeck. Carbon, nitrogen, and phosphorus transport by world rivers. *Am. J. Sci.*, 282(4):401–450, apr 1982.

[35] Markus Neteler, Hamish Bowman, Martin Landa, and Markus Metz. Grass gis: A multi-purpose open source gis. *Environmental Modelling & Software*, 31:124–130, 2012.

[36] Markus Neteler and Helena Mitasova. *Open source GIS: a GRASS GIS approach*, volume 689. Springer Science & Business Media, 2013.

[37] Kazuya Nishina, Akihiko Ito, Naota Hanasaki, and Seiji Hayashi. Reconstruction of spatially detailed global map of nh 4+ and no 3-application in synthetic nitrogen fertilizer. *Earth System Science Data*, 9(1), 2017.

[38] Josep Penuelas, Benjamin Poulter, Jordi Sardans, Philippe Ciais, Marijn Van Der Velde, Laurent Bopp, Olivier Boucher, Yves Godderis, Philippe Hinsinger, Joan Llusia, et al. Human-induced nitrogen–phosphorus imbalances alter natural and managed ecosystems across the globe. *Nature communications*, 4:2934, 2013.

[39] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2017.

[40] Nancy N Rabalais. Eutrophication of estuarine and coastal ecosystems. *Environmental Microbiology*, pages 115–134, 2010.

[41] Nn Rabalais, Re Turner, and Quay Dortch. Nutrient-enhanced productivity in the northern Gulf of Mexico: past, present and future. *Hydrobiologia*, pages 39–63, 2002.

[42] Timothy P Robinson, GR William Wint, Giulia Conchedda, Thomas P Van Boeckel, Valentina Ercoli, Elisa Palamara, Giuseppina Cinardi, Laura D'Aietti, Simon I Hay, and Marius Gilbert. Mapping the global distribution of livestock. *PloS one*, 9(5):e96084, 2014.

[43] M. Rodríguez Arredondo, P. Kuntke, A. W. Jeremiasse, T. H.J.A. Sleutels, C. J.N. Buisman, and A. Ter Heijne. Bioelectrochemical systems for nitrogen removal and recovery from wastewater. *Environ. Sci. Water. Res. Technol.*, 1(1):22–33, 2015.

[44] Helen Rowe, Paul JA Withers, Peter Baas, Neng Iong Chan, Donnacha Doody, Jeff Holiman, Brent Jacobs, Haigang Li, Graham K MacDonald, Richard McDowell, et al. Integrating legacy soil phosphorus into sustainable nutrient management strategies for future food, bioenergy and water security. *Nutrient Cycling in Agroecosystems*, 104(3):393–412, 2016.

[45] J.M. Sánchez-Pérez, F.A. Comín, S. Sauvage, J.J. Jiménez, and R. Sorando. Water resources and nitrate discharges in relation to agricultural land uses in an intensively irrigated watershed. *Sci. Tot. Environ.*, 659:1293–1306, 2018.

[46] C. Santhi, Jeffrey G. Arnold, J. R. Williams, W. a. Dugas, R. Srinivasan, and L. M. Hauck. Validation of the SWAT model on a large river basin with point and nonpoint sources. *J. Am. Water. Resour. Assoc.*, 37(5):1169–1188, 2002.

[47] S. P. Seitzinger, J. A. Harrison, Egon Dumont, Arthur H W Beusen, and A. F. Bouwman. Sources and delivery of carbon, nitrogen, and phosphorus to the coastal zone: An overview of Global Nutrient Export from Watersheds (NEWS) models and their application. *Global Biogeochem Cycles*, 19(4):1–11, 2005.

[48] Sukalyan Sengupta, Tabish Nawaz, and Jeffrey Beaudry. Nitrogen and Phosphorus Recovery from Wastewater. *Curr. Pollution. Rep.*, 1(3):155–166, 2015.

[49] Richard A. Smith, Richard B. Alexander, and Gregory E. Schwarz. Natural background concentrations of nutrients in streams and rivers of the conterminous United States. *Environment. Sci. Technol*, 37(14):3039–3047, 2003.

[50] Val Smith. Eutrophication of freshwater and coastal marine ecosystems a global problem. *Environment. Sci. Pollut. Res.*, 10(2):126–139, 2003.

[51] M.-N. Tuanmu and W. Jetz. A global 1-km consensus land-cover product for biodiversity and ecosystem modelling. *Global Ecology and Biogeography*, 23(9):1031–1045, 2014.

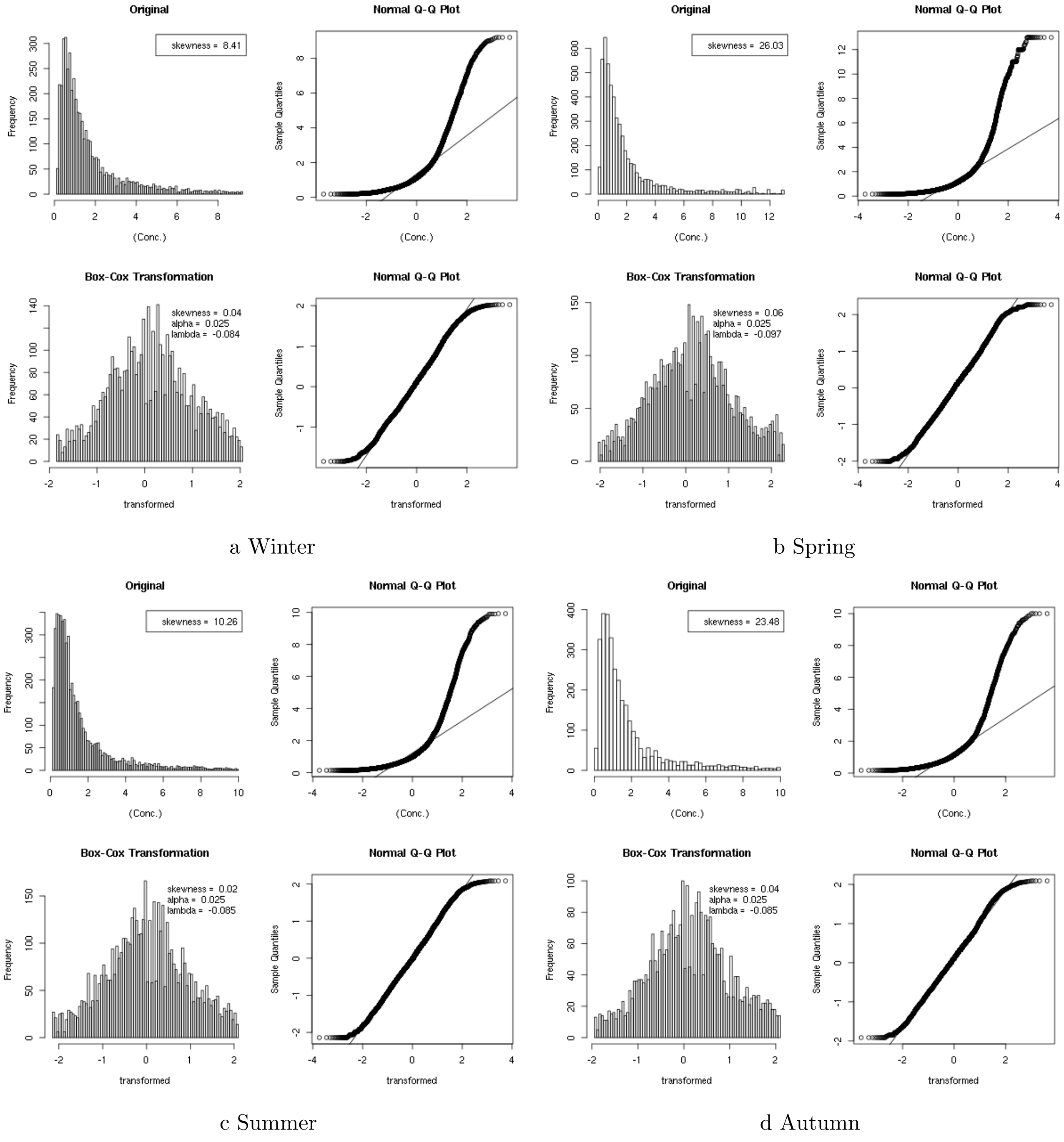
[52] Frank Warmerdam. The geospatial data abstraction library. In *Open source approaches in spatial data handling*, pages 87–104. Springer, 2008.

[53] P. G. Whitehead, E. J. Wilson, and D. Butterfield. A semi-distributed Integrated Nitrogen model for multiple source assessment in Catchments (INCA): Part I - Model structure and process equations, 1998.

[54] Dai Yamazaki, Daiki Ikeshima, Ryunosuke Tawatari, Tomohiro Yamaguchi, Fiachra O'Loughlin, Jeffery C Neal, Christopher C Sampson, Shinjiro Kanae, and Paul D Bates. A high-accuracy map of global terrain elevations. *Geophysical Research Letters*, 44(11):5844–5853, 2017.

A Supplementary information

Figure A.1: TN concentration (ppm) distribution plots for original value and Box Cox transformed values.



a Winter

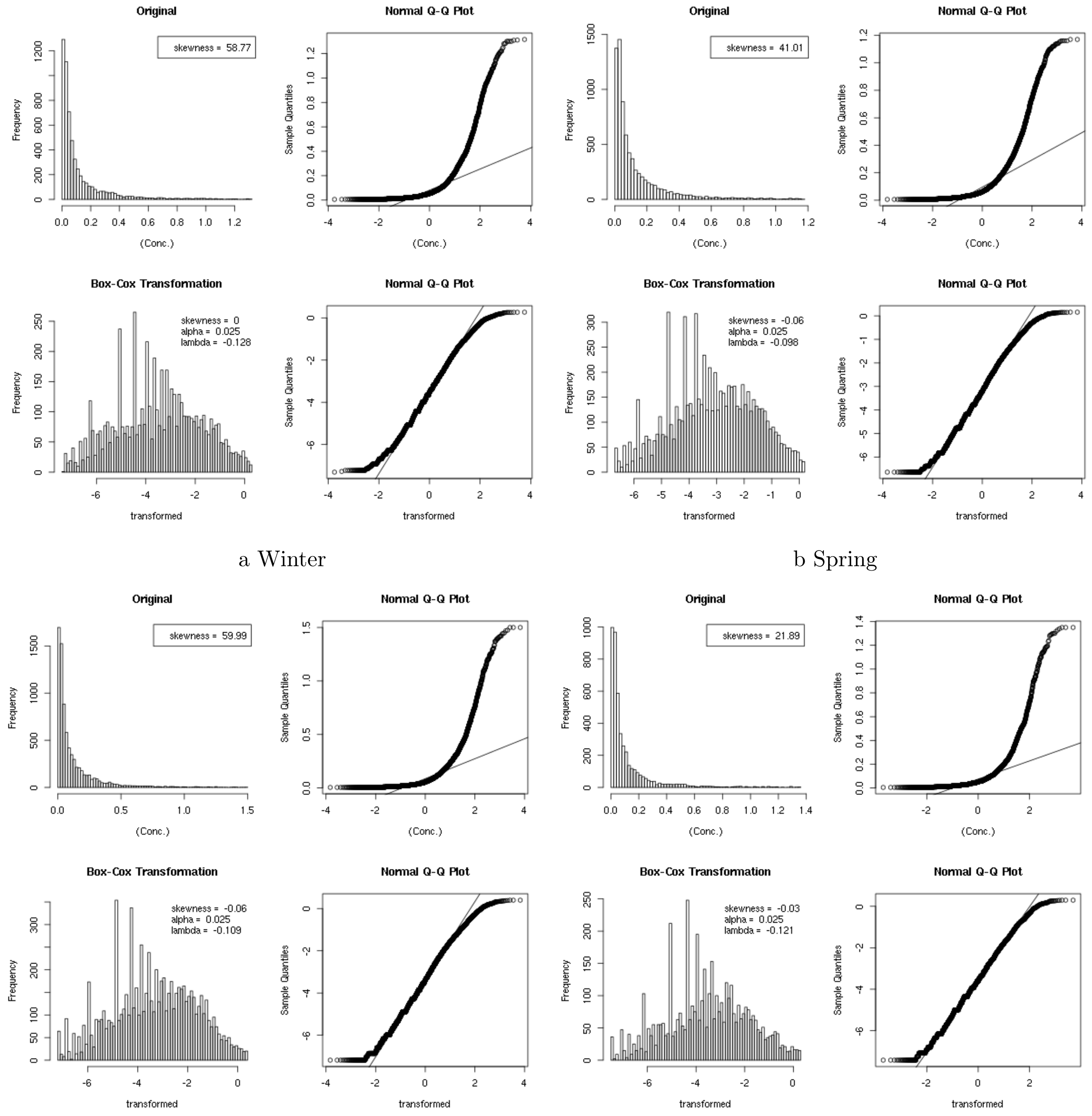
b Spring

c Summer

d Autumn

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27585v1 | CC BY 4.0 Open Access | rec: 13 Mar 2019, publ: 13 Mar 2019

Figure A.2: TP concentration (ppm) distribution plots for original value and Box Cox transformed values.



a Winter

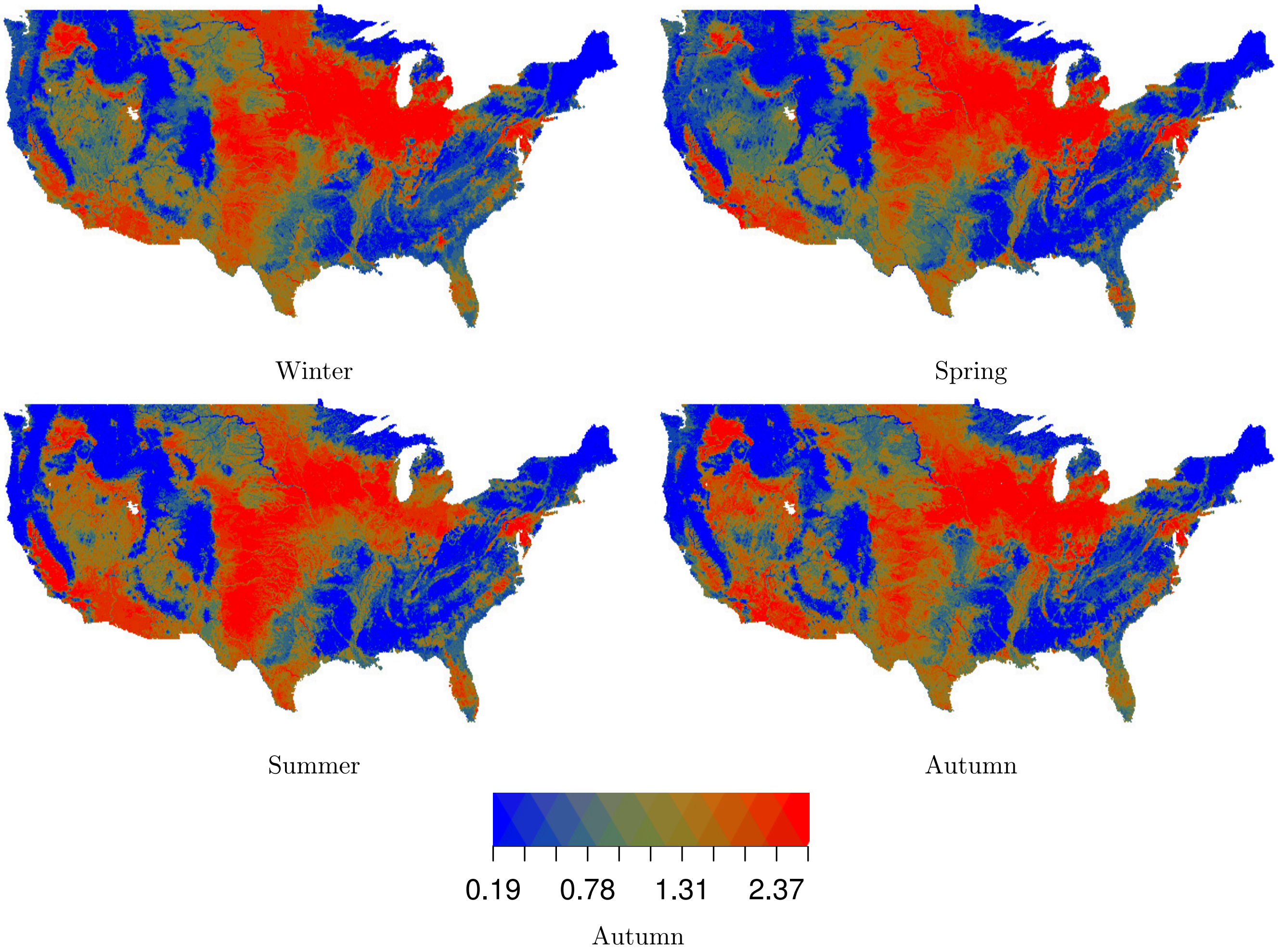
b Spring

c Summer

d Autumn

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27585v1 | CC BY 4.0 Open Access | rec: 13 Mar 2019, publ: 13 Mar 2019

Figure A.3: Seasonal prediction maps for TDN (ppm).



Winter

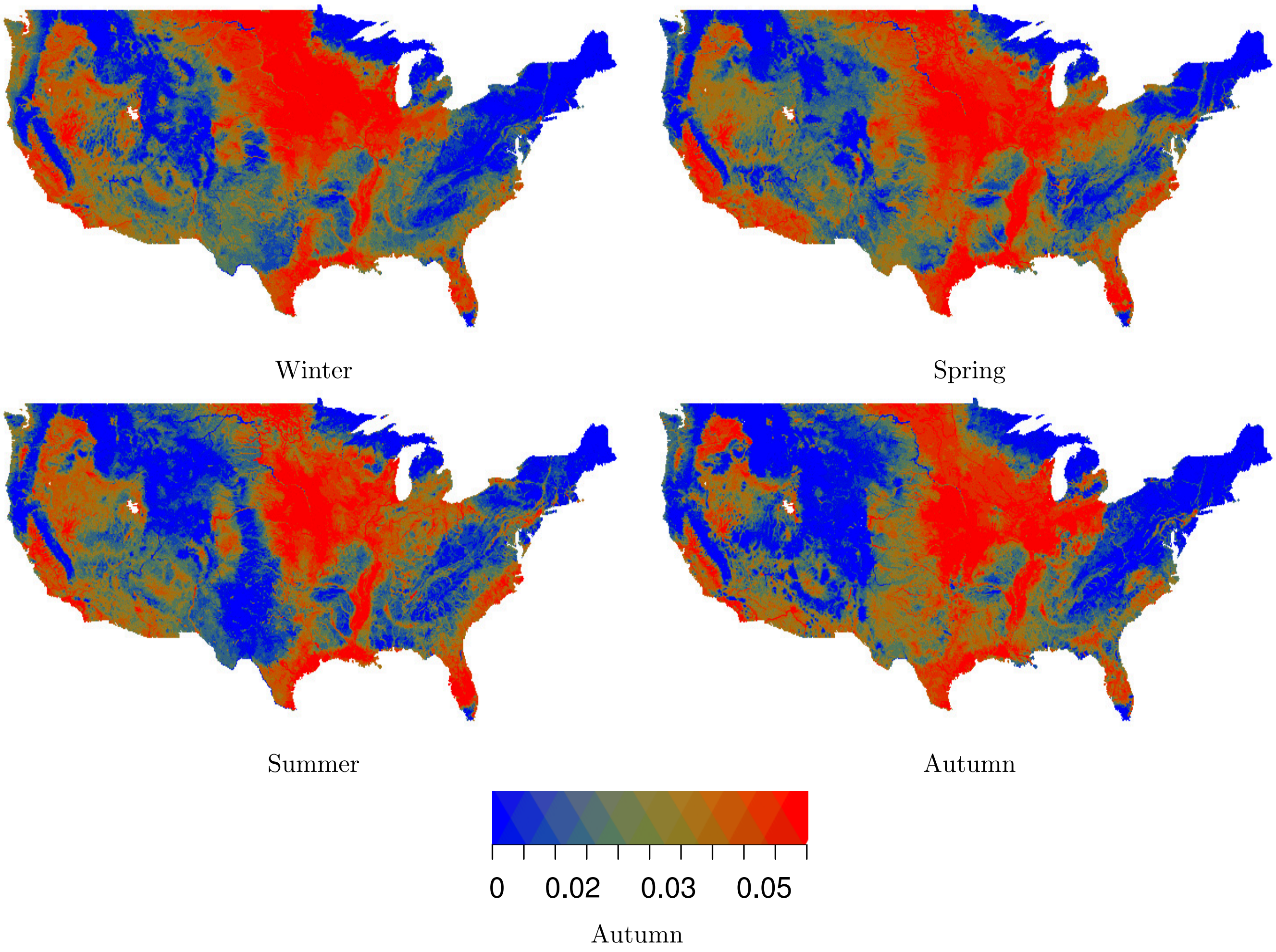
Spring

Summer

Autumn

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27585v1 | CC BY 4.0 Open Access | rec: 13 Mar 2019, publ: 13 Mar 2019

Figure A.4: Seasonal prediction maps for TDP (ppm).



Winter

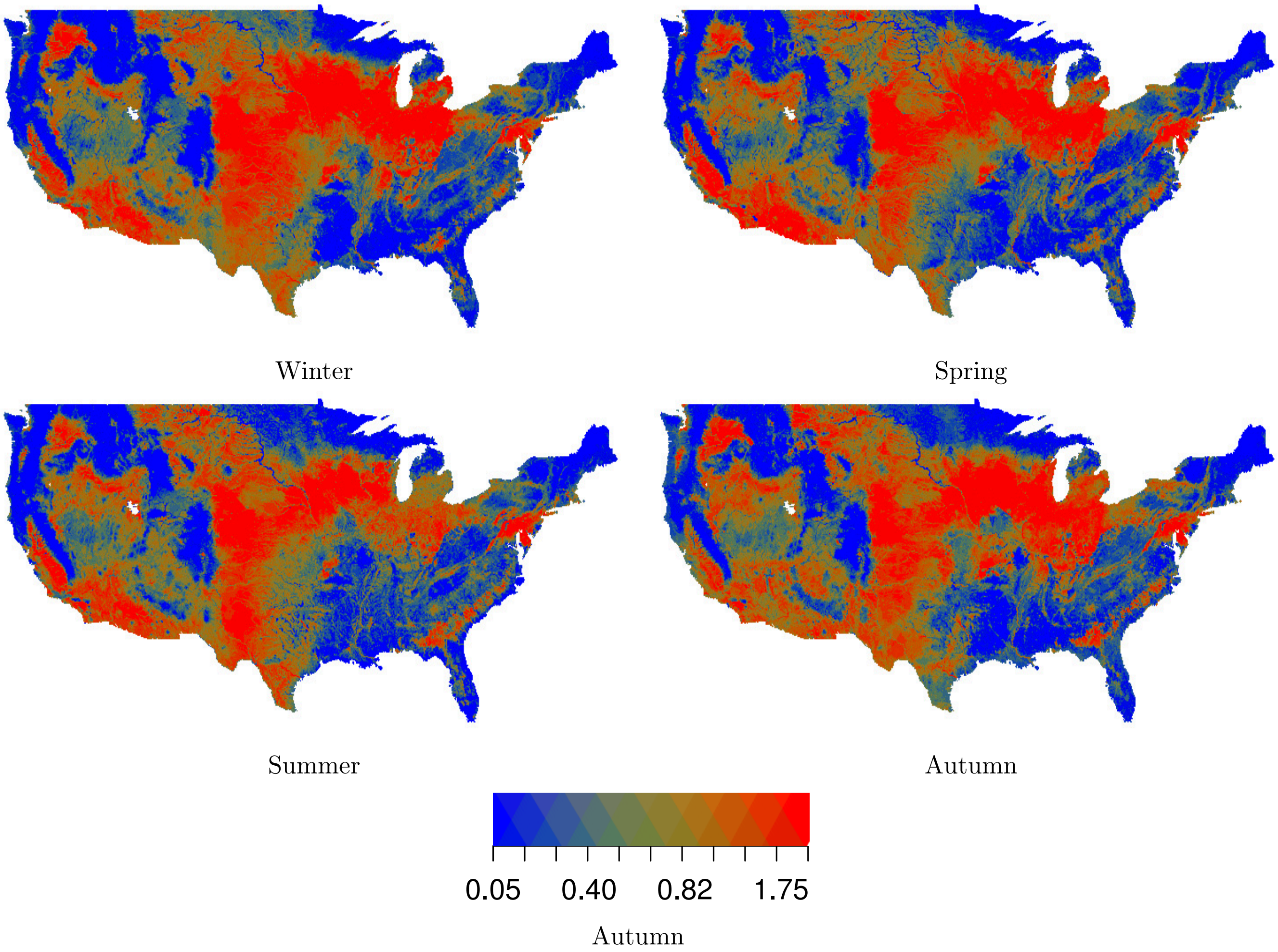
Spring

Summer

Autumn

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27585v1 | CC BY 4.0 Open Access | rec: 13 Mar 2019, publ: 13 Mar 2019

Figure A.5: Seasonal prediction maps for NO3 (ppm).



Winter

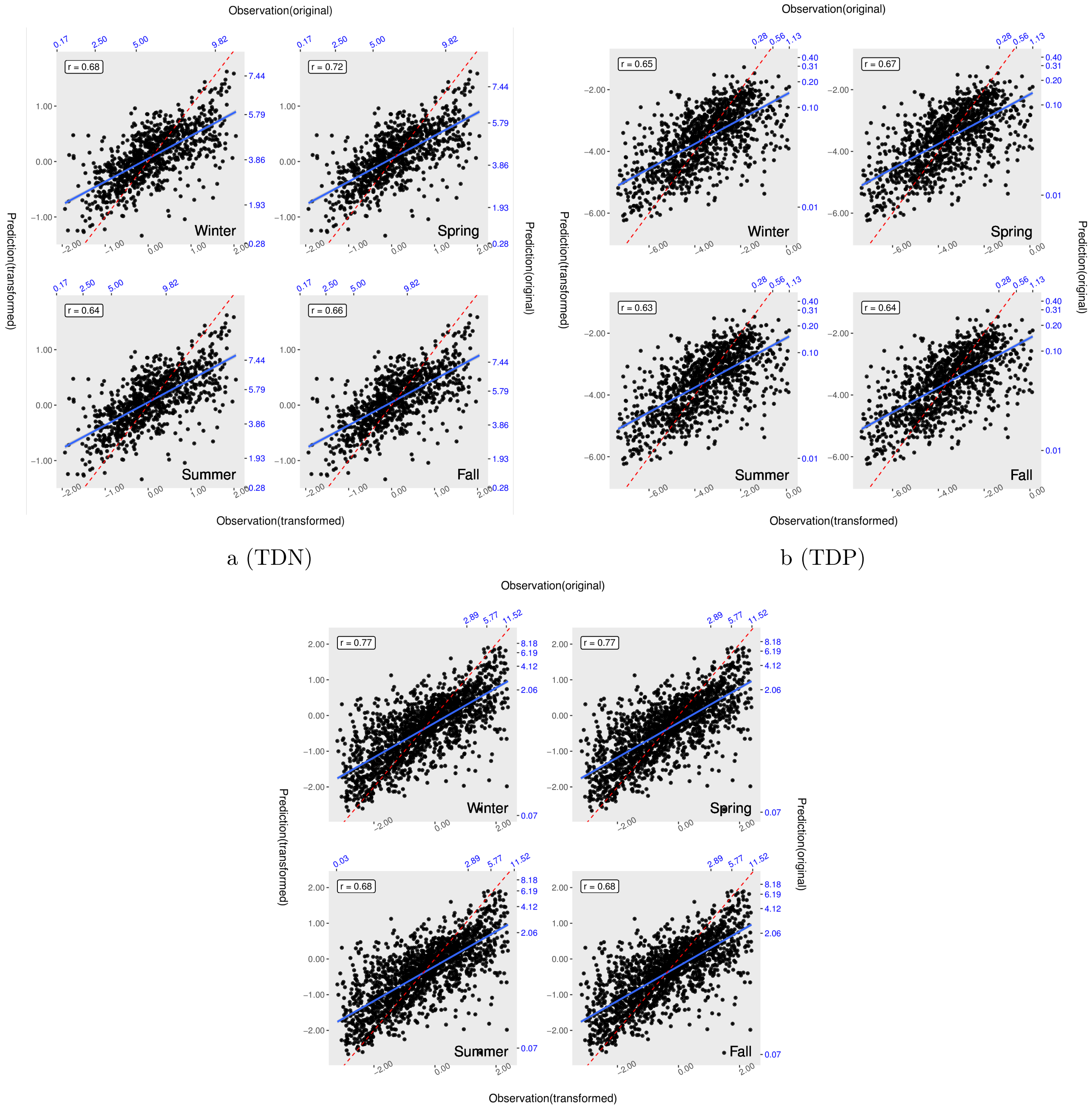
Spring

Summer

Autumn

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27585v1 | CC BY 4.0 Open Access | rec: 13 Mar 2019, publ: 13 Mar 2019

Figure A.6: Scatter plots regarding the model performance (testing dataset) for a) TDN, b) TDP and c) NO3, where each figure in the panel corresponds to a season.



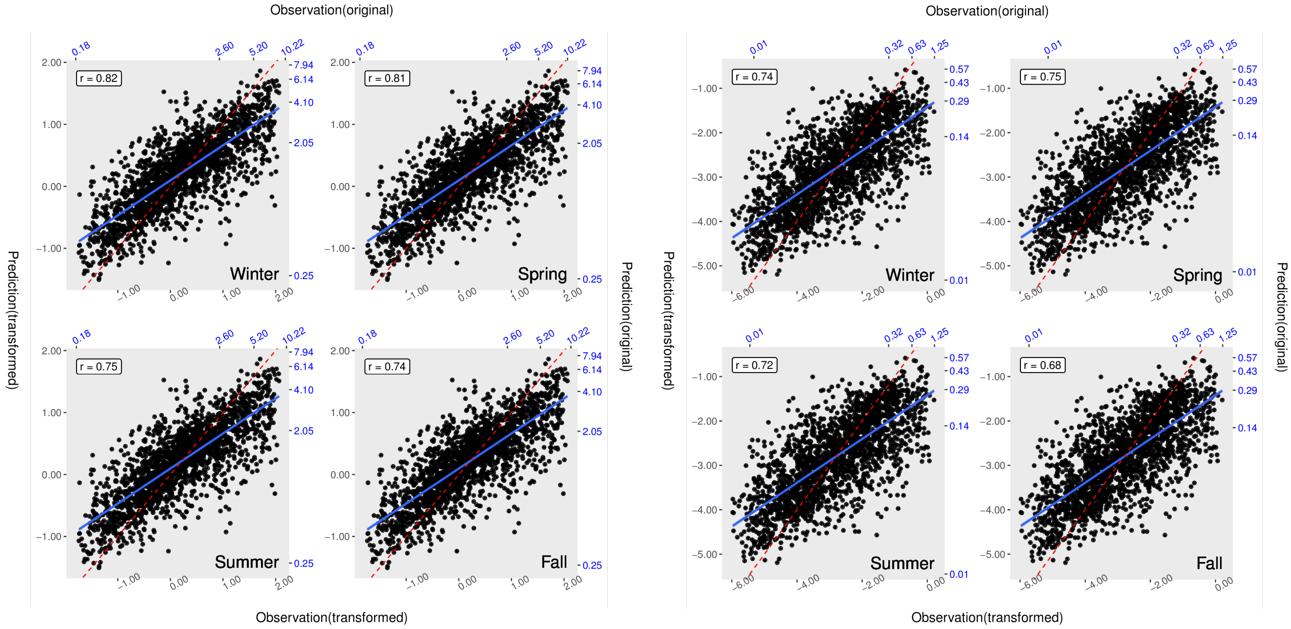
a (TDN)

b (TDP)

c (NO3)

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27585v1 | CC BY 4.0 Open Access | rec: 13 Mar 2019, publ: 13 Mar 2019

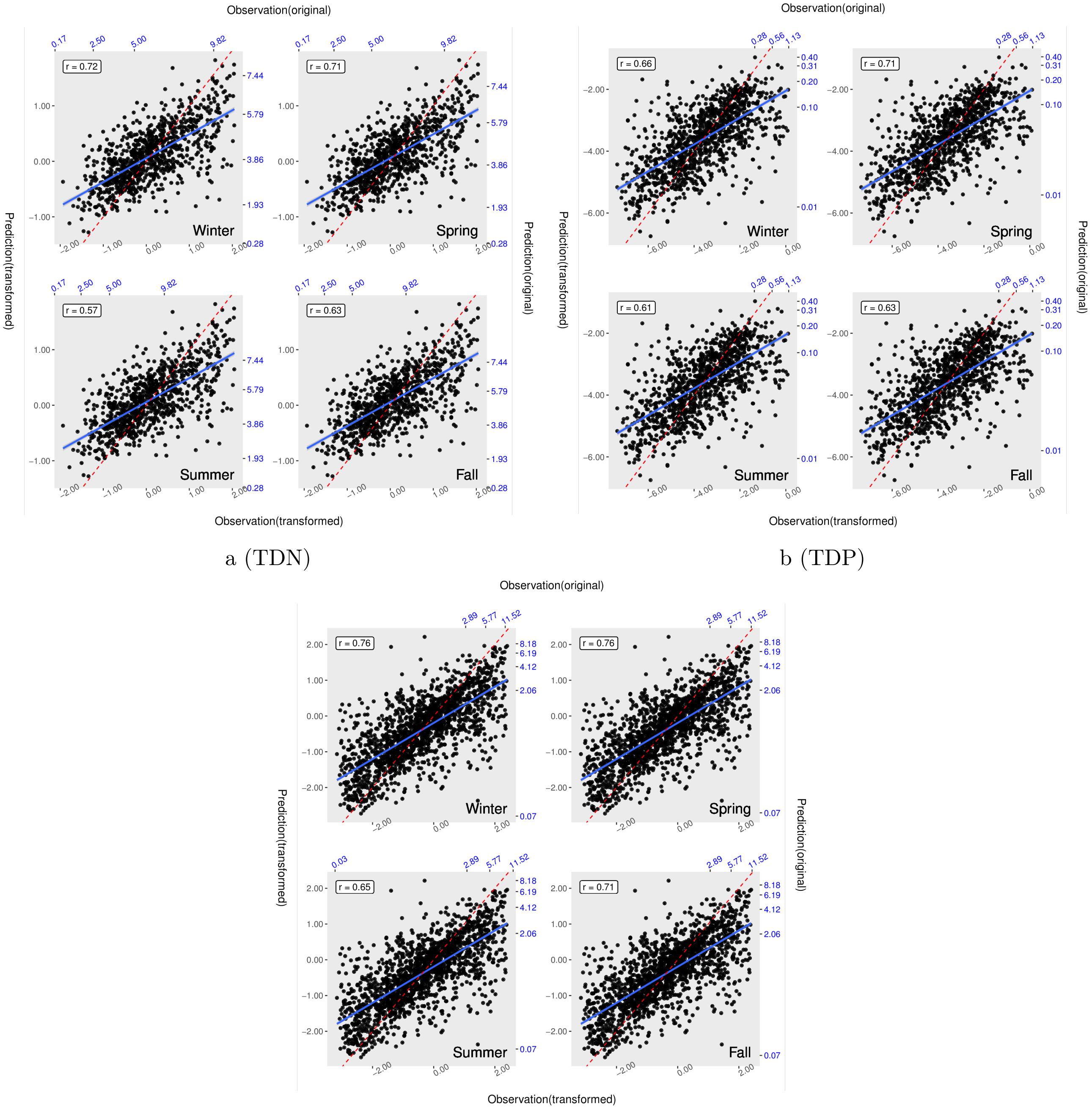
Figure A.7: Scatter plots regarding the model performance (internal validation - training) for a) TN, b) TP, where each figure in the panel corresponds to a season.



a (TN)

b (TP)

Figure A.8: Scatter plots regarding the model performance (internal validation - training) for a) TDN, b) TDP and c) NO3, where each figure in the panel corresponds to a season.

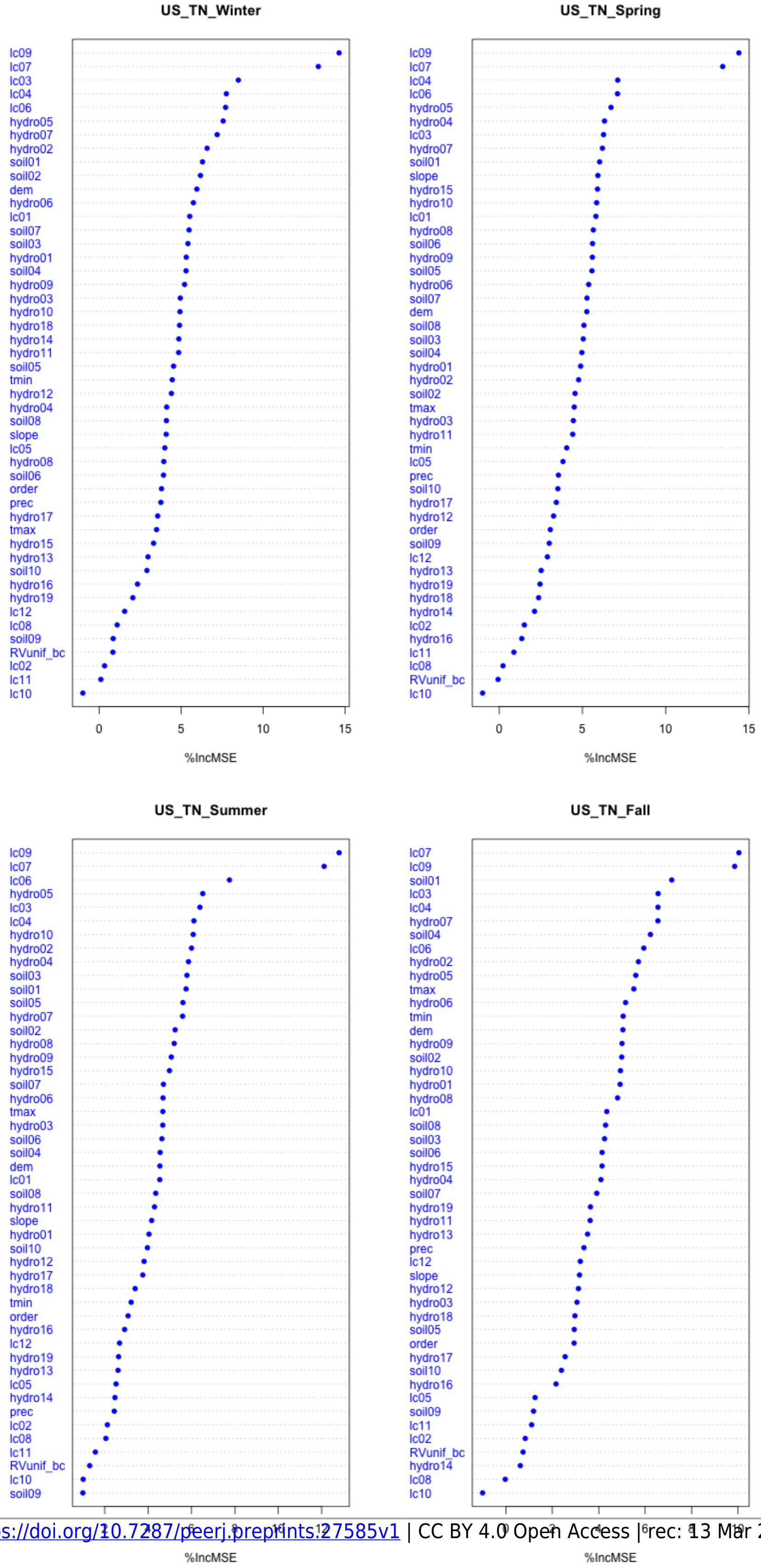


a (TDN)

b (TDP)

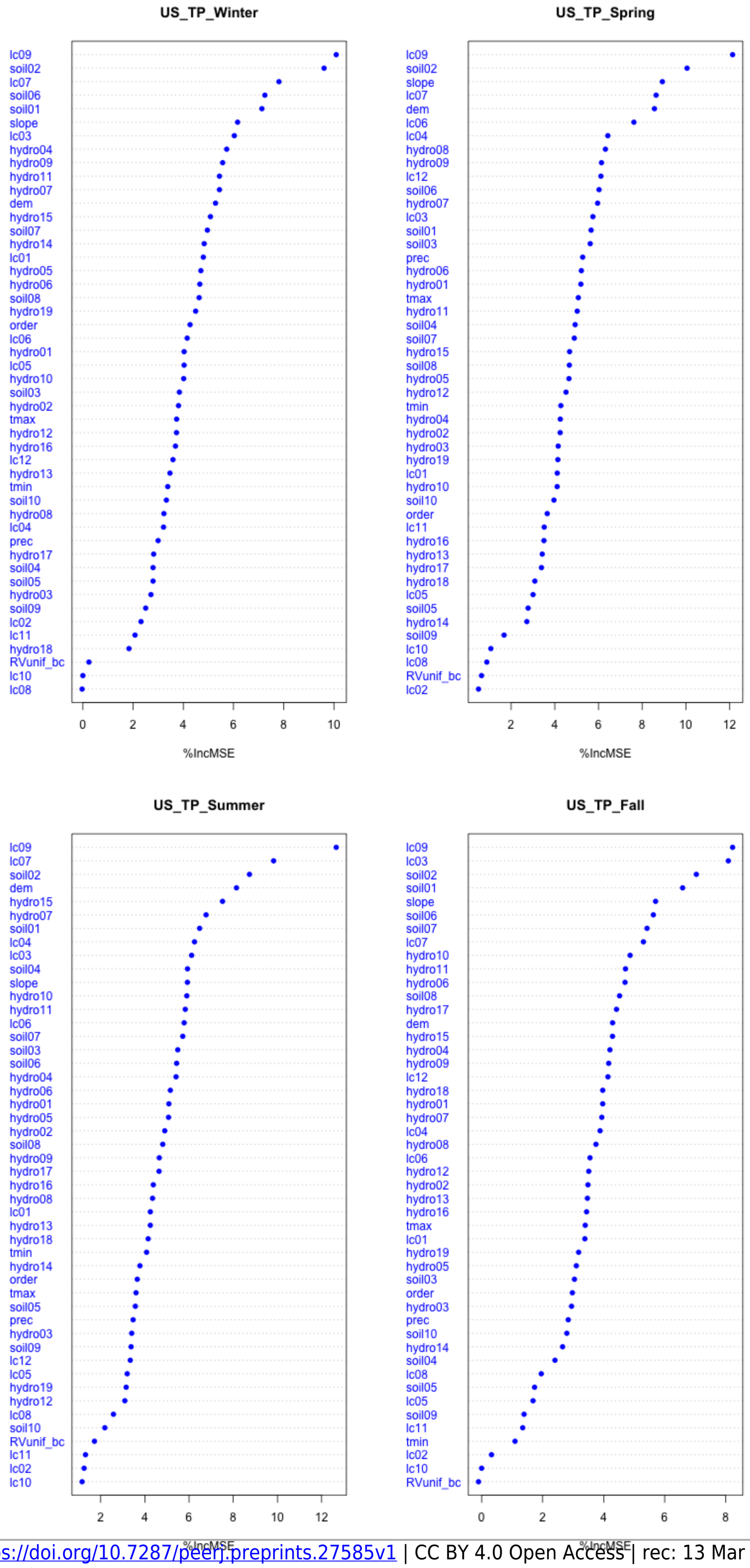
c (NO3)

Figure A.9: Variable importance expressed as percent contribution of the Mean Square Error for each predictor in the model for TN across the four seasons. The most important variables are Urban/Built-up (lc09) and Cultivated and Managed Vegetation (lc07). RVuni_bc stands for a random variable drawn from a uniform distribution to access the relative importance of other physically derived variables.



(TN)

Figure A.10: Variable importance expressed as percent contribution of the Mean Square Error for each predictor in the model for TP across the four seasons. The first most important variable is Urban/Built-up (lc09), rather the second one can be Cultivated and Managed Vegetation (lc07) or Soil pH in H2O (soil02).



(TP)

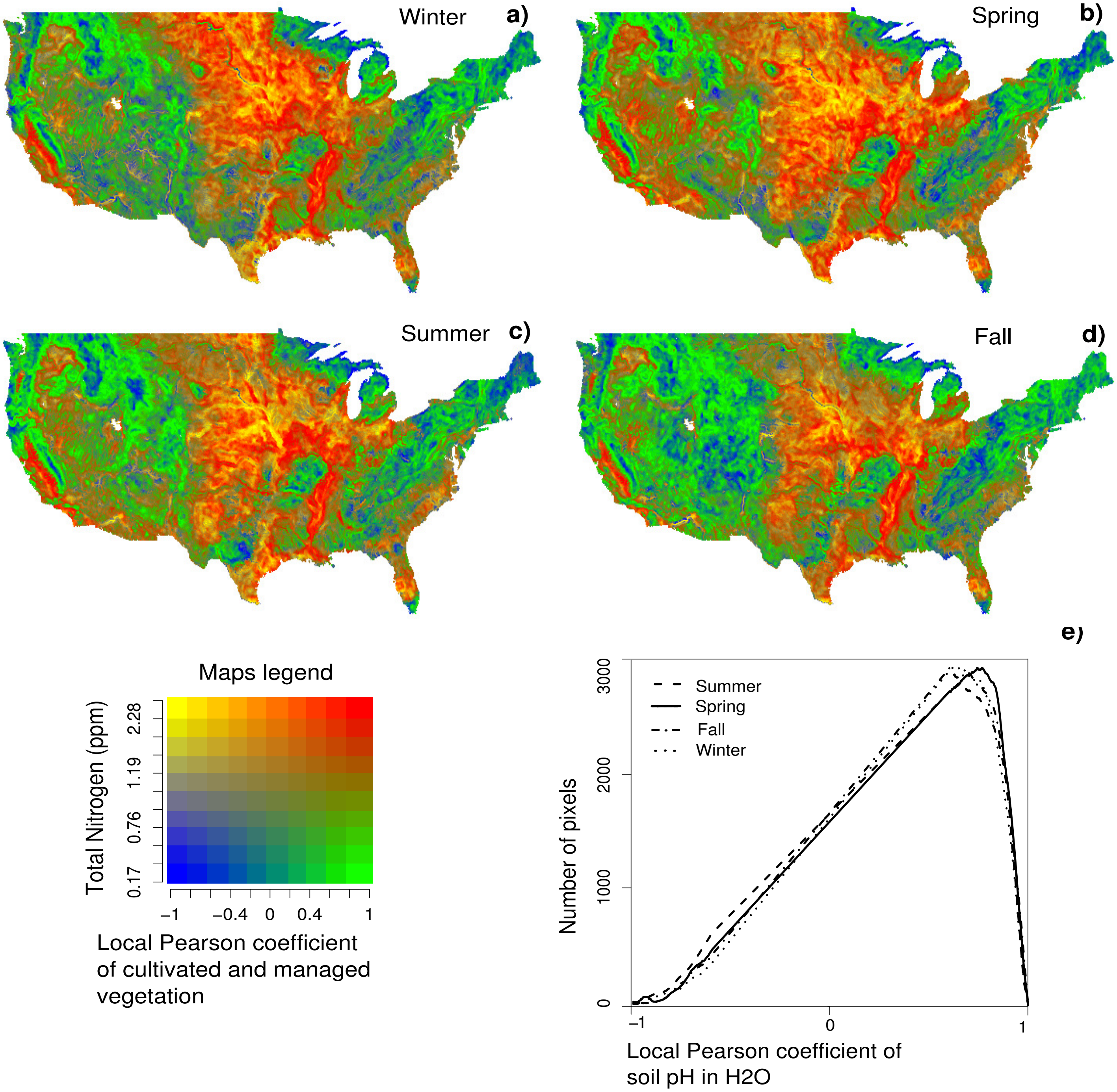


Figure A.11: a-d) Bivariate maps representing the relationship between the Total Phosphorus (TP) and the local Pearson coefficient of soil pH in H2O landcover vs TP. e) Graph showing the histogram tails of the raster map of the local Pearson coefficient. The maps is able to show the influence/correlation of the urban/built-up (expressed in percentage) to the level of TP. A positive high correlation could appear in areas with low level of TP (green), or in areas with high TP (red). Similarly, negative correlation could occur in low concentration areas(blue) or in high concentration areas (yellow). Streams and rivers are aggregated by a moving window of 10x10 grid-cells depicting the the mean value for a better visualisation effect. The graph highlights a stronger positive correlation of the urban/built-up during warmer seasons (summer and spring) in comparison with the colder seasons (autumn and winter).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27585v1 | CC BY 4.0 Open Access | rec: 13 Mar 2019, publ: 13 Mar 2019

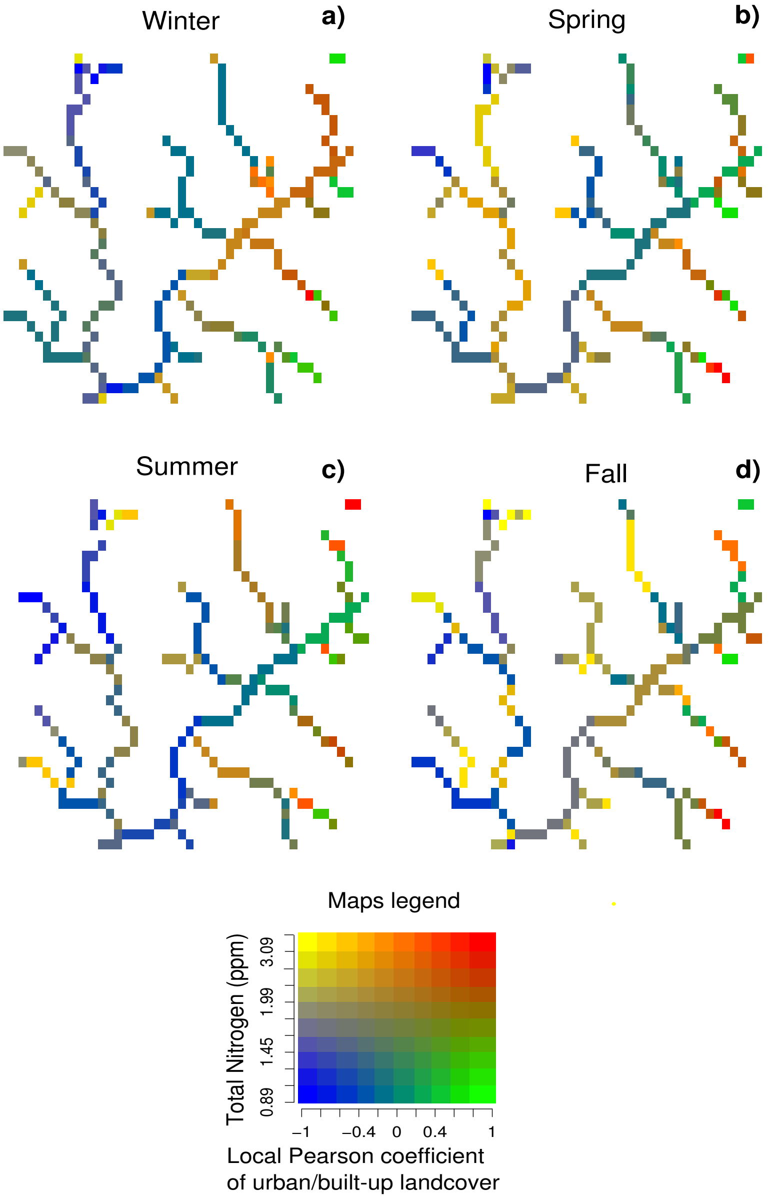


Figure A.12: a-d) Bivariate maps representing the relationship between the Total Nitrogen (TN) and the local Pearson coefficient of urban/built-up landcover vs TN at stream level. High variability of correlation at cell level is evident, denoting the capability of RF picking up spatial variability.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27585v1 | CC BY 4.0 Open Access | rec: 13 Mar 2019, publ: 13 Mar 2019