Carbon dioxide partial pressure and emission throughout the Scandinavian stream network

Kenneth Thorø Martinsen[1], Theis Kragh[2] and Kaj Sand-Jensen[1]

[1]Freshwater Biological Laboratory, Biological Institute, University of Copenhagen, Universitetsparken 4, 3rd. floor, DK-2100 Copenhagen Ø, Denmark

[2]Biological Institute, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark

Corresponding author: Kenneth Thorø Martinsen (kenneth2810@gmail.com)

Key Points:

- Carbon dioxide partial pressure in streams can be predicted from catchment characteristics
- Prediction of carbon dioxide partial pressure can be improved using machine learning
- Prediction of carbon dioxide partial pressure, followed by estimation of fluxes, can be performed across large stream networks

Martinsen, K. T., Kragh, T., & Sand-Jensen, K. (2020). Carbon dioxide partial pressure and emission throughout the Scandinavian stream network. Global Biogeochemical Cycles, 34, e2020GB006703. https://doi.org/10.1029/2020GB006703

Abstract

Stream networks transport and emit substantial volumes of carbon dioxide (CO_2) into the atmosphere. We gathered open monitoring data from streams in three Scandinavian countries and estimated CO_2 partial pressure (pCO_2) at 2298 sites. Most of the sites (87 %) were supersaturated when averaged across the year with an overall mean pCO_2 of 1464 μatm (range: 17–15646). Using remote sensing data, we modeled a realistic stream network including streams above ~2.5 m wide and calculated catchment averages of multiple variables associated with geo-morphometry, stream network proximity and land cover. We compared the ability of eight machine learning models to predict pCO_2 and found that the Random Forest model achieved the highest accuracy, with a root-mean-square error of 0.22 ($\log_{10}(pCO_2)$) and R^2 of 0.66. Mean catchment elevation, slope and permanent water cover were the most important predictor variables. We used the predictive model to create a high-resolution (25 m resolution) map with predicted stream pCO_2 throughout the 268.807 km stream network in Denmark, Sweden and Finland. Predicted pCO_2 averaged 1134 μatm (range: 154–8174). We used surface runoff, air temperature and stream channel slope to estimate gas transfer velocity and CO_2 flux throughout the network. Mean stream CO_2 fluxes ranged from 1.0 and 1.2 in Sweden and Finland respectively 3 to 3.2 g C m^{-2} d^{-1} in Denmark. Better-performing models improve our ability to predict pCO_2 in stream networks and reduce the uncertainty of upscaling estimates of carbon emissions from inland waters to countries and continents.

1 Introduction

Inland standing and running waters are hotspots of greenhouse gas emissions on both regional and global scales (Battin et al., 2009; Raymond et al., 2013). Knowledge of the magnitude of these emissions has grown in recent years, as efforts to reduce the uncertainty of the global freshwater carbon emissions have increased (Drake et al., 2018). In terms of both carbon dioxide (CO_2) emissions and lateral transport, headwaters, streams and rivers (simplified to ‘streams’ herein) play a dominant role (Butman et al., 2015; Wallin et al., 2013); relative to the area they cover, CO_2 emissions from streams are disproportionately high (Butman & Raymond, 2011; Marx et al., 2017). Improving the predictive accuracy of CO_2 partial pressure (pCO_2) in streams is important for constraining the role of stream CO_2 flux (F_{CO2}) in large scale greenhouse gas emission budgets.

The earlier view of streams acting only as “pipes” or closed transport ways of carbon from the terrestrial environment to coastal areas has long been outdated (Cole et al., 2007). Instead, streams are understood not only to transport carbon, but also to emit it into the atmosphere. Processing of allochthonous organic carbon in the aquatic environment and input of terrestrially derived CO_2 result in pronounced supersaturation (Humborg et al., 2010). This supersaturation results in diffusive transport of CO_2 from water to the atmosphere. The rate of this transport (the flux) depends on the gas transfer velocity (*k)* and the difference between the actual and saturation partial pressure (concentration being the product of Henry’s law constant, H_p, and the partial pressure):

Eq. 1: $F_{CO_2}=k\cdot H_p\cdot\left(pCO_2-pCO_{2sat}\right)$

k depends on surface water turbulence and can be estimated from physical characteristics such as stream hydraulics (Zappa et al., 2007). Recent large-scale studies have applied empirical relationships, scaling *k* by water velocity and stream slope (Raymond et al., 2012). Thus, F_{CO2} can be calculated with knowledge of pCO_2 in the water, water temperature and estimates of *k* from water velocity and stream slope.

Streams connect the terrestrial environment to the ocean. Water is received along the stream network from a catchment via groundwater, soil water and surface water discharge. From the moment carbon enters a stream, it may be transformed and emitted to the atmosphere as CO_2 or transported further downstream. Both inorganic and organic carbon may originate from aquatic metabolic processes (Sand-Jensen et al., 2007), photodegradation (Cory et al., 2015) and hydrological inputs from the catchment (Humborg et al., 2010). CO_2 derived from soil respiration contributes markedly to stream pCO_2 (J. B. Jones & Mulholland, 1998a). This may be especially pronounced in headwaters, where 50–100-fold supersaturation has been documented (Johnson et al., 2008), and similarly high pCO_2 values have been observed in agricultural lowlands (Sand-Jensen & Staehr, 2012). Depending on the catchment geology, much of CO_2 in the soil is converted to bicarbonate ions by chemical weathering of carbonates and aluminum-silicate minerals (Berner & Berner, 2012; Marx et al., 2017). Catchment geo-morphometry, land cover and stream network proximity could provide useful proxies for hydrological carbon loading and should improve the predictability of stream pCO_2.

Estimates of F_{CO2} on larger scales are often based on averages of CO_2 concentration and *k* for different stream order categories, thus ignoring spatial variability, which is likely to be influential. However, recent studies (Horgby et al., 2019; Rocher-Ros et al., 2019) show the importance of catchment scale processes on pCO_2 in stream networks. It is thus attractive to develop methods that enable easier incorporation of such processes into large-scale budgets to improve their accuracy. The use of data-driven modeling focused on prediction and leveraging geographical information systems (GIS) for the extraction of relevant drivers could provide such a framework. Multiple studies have identified average catchment slope as an important predictor of pCO_2 levels (Hutchins et al., 2019; Lauerwald et al., 2015; Smits et al., 2017) which, along with climate variables such as temperature and precipitation, influence soil respiration, water status and hydrological connectivity between catchment and stream. Furthermore, land use and land cover, which can be obtained by remote sensing, could also prove to be valuable. The presence of more upstream lakes is expected to result in lower downstream pCO_2 levels (Sand-Jensen & Staehr, 2012) while CO_2 production and hydrological connectivity might increase during periods with ephemeral water cover (Johnson et al., 2008; Marcé et al., 2019) as also evident in temporary wetlands (Abril et al., 2014). Higher pCO_2 levels in agricultural catchments compared to forest-dominated catchments, have also been observed (Borges et al., 2018). Finally, the influence on pCO_2 of catchment input relative to in-stream processes is expected to decrease with stream width and, thus, with distance from the source (Hotchkiss et al., 2015). This gradual shift in the contribution of in-stream and catchment processes to pCO_2 production is accompanied by changing downstream hydrology as the catchment area and stream discharge increase. For example, S. Liu & Raymond (2018) show that the relationship between pCO_2 and stream discharge was negative in small streams and positive in large streams, but the F_{CO2}–stream discharge relationship was generally positive across all stream orders. Therefore, stream discharge plays a dual role in pCO_2–F_{CO2} dynamics, by controlling stream hydraulics and in turn *k*, the CO_2 input from the catchment and the in-stream CO_2 generation. Consequently, identifying important climate and environmental drivers of pCO_2 is also the key to improving the accuracy of upscaled estimates of F_{CO2} from stream sites to fluvial networks. Additionally, if data related to the important variables are readily available, then predictions of large-scale carbon emissions can be made regarding sparsely investigated or remote areas.

Previous studies have often applied linear models and stepwise model selection (Horgby et al., 2019; Hutchins et al., 2020; Weyhenmeyer et al., 2012) to approximate the relationship between stream pCO_2 and its predictors. However, the proportion of explained variation has generally been low. An attractive approach for improving our ability to predict pCO_2 in stream networks is to apply methods from the field of machine learning. These techniques have been increasingly applied in ecological, hydrological and biogeochemical studies (Barbarossa et al., 2018; Olden et al., 2008). Machine learning offers several efficient models that may achieve high predictive accuracy of new observations (James et al., 2013). As opposed to more traditional statistical techniques such as a linear regression model, the functional relationship between response and predictors are not defined beforehand but instead “learned” from data. The flexibility of the models is tuned based on their ability to generalize on new observations. Not only does this way of fitting models to data result in more accurate predictions, but new insights on important drivers can also emerge. These models can be applied to areas of interest because by incorporating the spatial drivers explicitly using geographical information systems (GIS), we can obtain high-resolution maps of pCO_2 in streams. High resolution is crucial for determining the role of streams in large-scale carbon emissions and at the same time is useful for catchment-scale studies. Reducing the uncertainty associated with predicting pCO_2 at new sites improves the accuracy of up-scaled emission estimates and, in turn, our ability to constrain the inland waters in global scale carbon emissions.

In an attempt to improve our understanding of carbon dynamics and emissions in large scale river networks we compiled open geospatial and water chemistry data and applied machine learning methods to predict pCO_2. Scandinavia spans large gradients in geology, climate and land cover, and it is representative of the north temperate zone, making it a suitable study region to explore our suggested approach. Therefore, we compiled open data on alkalinity, pH and water temperature and estimated pCO_2 from stream sites in Denmark, Sweden and Finland. We derived catchment and climate characteristics believed to be important predictors of pCO_2. We compared multiple predictive models and combined the best one with a high-resolution model of the stream network, in order to predict pCO_2 levels throughout the network. To demonstrate how this could be used in large scale carbon budgets, we also estimated the mean annual *k* and F_{CO2} using a coarser-resolution map of surface runoff. Our hypothesis: pCO_2 in stream networks can be accurately predicted from catchment characteristics and form the basis for future estimates of F_{CO2}.

2 Methods

2.1 Estimation of pCO_2

We downloaded open monitoring data from environmental agencies in Denmark (DNK), Sweden (SWE) and Finland (FIN; MFVM & DCE, 2019; MVM, 2019; SYKE, 2019). We selected observations where data on water temperature, pH, alkalinity and coordinates were available for the period 1990–2018. Data on water depth were also downloaded when available; samples taken from deeper than 2 meters were discarded to retain surface water observations only. Estimation of pCO_2 from pH and alkalinity is potentially biased due to the possible influence of non-carbonate alkalinity (Abril et al., 2015), especially in low-alkaline regions. To minimize alkalinity- and pH-related biases in the calculation of pCO_2, we applied the corrections for pH measurement error and organic alkalinity described in Liu et al. (2020) and only used alkalinity values between 0 and 10 meq L^{-1}. We calculated pCO_2 using the *seacarb* R-package (Gattuso et al., 2018) and excluded observations when the calculated pCO_2 exceeded 40,000 μatm, as we believed these to be biased. In order to calculate a robust annual average, we used the same procedure as Lauerwald et al. (2015). First, we calculated the monthly median pCO_2 for each site discarding months with less than three observations. Secondly, we performed linear interpolation between months discarding sites with gaps exceeding three months. Finally, we calculated the annual mean pCO_2 for a total of 2298 sites.

2.2 Spatial data processing

We downloaded open, remote-sensing data products made available by the EU (EEA, 2016) and previous studies and derived several variables that are likely to influence stream pCO_2. We used the WorldClim version 1.4 data product, which included mean annual air temperature and annual precipitation (Hijmans et al., 2005). These include variables related to geo-morphometry, climate, land cover and stream network proximity. More specifically, we used a digital elevation model (DEM, EU-DEM version 1.1), high-resolution themes (grassland, water, forest) and Corine Land Cover (agriculture) data products. We created binary layers of the categories included in each high-resolution theme, resulting in multiple presence/absence layers of grassland, coniferous forest, broad-leaved forest, permanent water, temporary water, permanent wetness and temporary wetness.

To ease computations, the region was split into five major basins, each with a buffer of 25 km to avoid edge effects. TauDEM software (version 5.3; Tarboton, 2017) was used for hydrological processing of DEMs and to calculate height above nearest drainage (HAND) and stream proximity metrics (e.g. stream order and network length). In order to create a realistic model of the stream network, we adapted the approach described by Y. Y. Liu et al. (2018). For this purpose, we used a stream map (EU-HYDRO) based on photo-interpretation of very-high-resolution imagery (EEA, 2017). This ensures consistency between the modeled stream network and calculated flow directions. Hydrological conditioning of DEMs enabled us to remove obstacles (e.g. roads and culverts) along the stream-lines and pits; Garousi-Nejad et al. (2019) describe the procedure in detail. Following DEM preprocessing, we calculated flow direction (deterministic-8), catchment area, stream order, total flow path and longest flow path. We calculated HAND along the entire network and used the proportion of the catchment area with a HAND less than 2 m as a proxy for stream–groundwater connectivity. We determined catchment averages for continuous variables and proportions for categorical predictor variables for all grid cells in the stream network. Sites not coinciding with the stream network were moved along the flow direction for a maximum distance of 250 m (10 times the resolution) towards the modeled stream network and sites with a drainage area below 6.25 ha (10*10 times the resolution) were discarded. We used a suite of open-source software for the spatial analysis (GDAL & OGR, 2018; Hijmans, 2019; McInerney & Kempeneers, 2015; Pebesma, 2018).

Martinsen, K. T., Kragh, T., & Sand-Jensen, K. (2020). Carbon dioxide partial pressure and emission throughout the Scandinavian stream network. Global Biogeochemical Cycles, 34, e2020GB006703. https://doi.org/10.1029/2020GB006703

2.3 Predictive modeling

To evaluate the performance of candidate models and assess the accuracy of the final model, respectively, 80 % of the initial 2293-observation dataset was designated as the training set; the remaining 20 % of observations comprised the test set. As the density distributions of many predictor variables were skewed, we applied preprocessing using the Yeo-Johnson power-transformation (Yeo & Johnson, 2000) followed by standardization (subtracting the mean and dividing by the standard deviation). To avoid intercorrelation between predictors, we excluded all variables (catchment proportion of agriculture, Strahler order, total and longest network path) with an r-value (Pearson correlation) greater than 0.65. This left 11 explanatory variables available for predictive modeling (Table 1). The response variable pCO_2 was $\log_{10}$ transformed to improve the normality of the distribution.

Table 1. Units, ranges, sources and resolutions (original/used) of catchment variables used for predictive modeling of stream pCO2.

Variable	Unit	Range	Resolution (m)	Data source
catch. avg. elevation	m	0–965	25/25	EU-DEM v1.1
catch. avg. slope	°	0.2–11.9	25/25	EU-DEM v1.1
catch. area	km^2	0–50032	25/25	EU-DEM v1.1
catch. prop. HAND < 2 m		0–0.9	25/25	EU-DEM v1.1
catch. prop. grassland		0–0.7	20/25	EU Grassland 2015
catch. prop. coniferous forest		0–0.9	20/25	EU Dominant leaf type 2012
catch. prop. broad-leaved forest		0–0.7	20/25	EU Dominant leaf type 2012
catch. prop. perm. water		0–0.5	20/25	EU Water and Wetness 2015
catch. prop. temp. water		0–0.1	20/25	EU Water and Wetness 2015
catch. prop. perm. wet		0–0.1	20/25	EU Water and Wetness 2015
catch. prop. temp. wet		0–0.6	20/25	EU Water and Wetness 2015
catch. avg. annual precipitation	mm	487–1142	~1000/25	Worldclim 1.4

In order to select the best-performing model to predict pCO_2, we compared several machine-learning models of different complexity. We compared their performance using 5-fold cross-validation repeated 25 times (outerloop for performance estimation) on the training set. For optimal performance, most predictive models depend on the tuning of hyperparameters. To estimate the optimal hyperparameters in our case, we defined probable search spaces and used random sampling (50 iterations) and 5-fold cross-validation (inner tuning loop). The performances of the trained models were compared by using the root-mean-square error (RMSE) and the coefficient of determination (R^2). The best-performing model was then tuned and trained on the training set. Here, the optimal hyperparameter settings were identified after a more exhaustive search using sequential model-based optimization with a maximum of 200 iterations (Bischl et al., 2017). With the training set and the same tuning procedure used for model comparison, we also assessed model performance when extrapolating to new geographical regions using a “leave-one-country-out” cross-validation scheme. Predictive modeling was performed using the machine-learning meta-package *mlr* in R (Table 2; Bischl et al., 2016).

Martinsen, K. T., Kragh, T., & Sand-Jensen, K. (2020). Carbon dioxide partial pressure and emission throughout the Scandinavian stream network. Global Biogeochemical Cycles, 34, e2020GB006703. https://doi.org/10.1029/2020GB006703

Table 2. Candidate models for prediction of pCO_2, their associated R-packages and the hyperparameters tuned during model training.

Model	Package	Hyperparameters
Linear Model	stats	
k-Nearest Neighbour	FNN	K
Decision Tree	rpart	cp, maxdepth, minbucket, minsplit
GLM with Elasticnet regularization	glmnet	alpha, lambda
Neural Network	nnet	size, decay
Support Vector Machine	kernlab	C, sigma
Random Forest	ranger	mtry, num.trees, sample.fraction, min.node.size
Extreme Gradient Boosting	xgboost	nrounds, max_depth, eta, subsample, min_child_weight, alpha, lambda

2.4 Estimating F_{CO2}

Annual mean runoff data from Beck et al. (2015) and empirical relationships were used to estimate annual mean *k* and subsequently F_{CO2} throughout the stream network. We accumulated runoff throughout the stream network to estimate annual mean discharge (Q, m^3 s^{-1}). Empirical relationships were used to estimate flow velocity (V, m s^{-1}) from discharge (Raymond et al., 2012) and k_{600} (m d^{-1}) from flow velocity and stream slope (S, m m^{-1}; eq. 5 in Table 2 in Raymond et al. (2012)). Stream length (L) was approximated as the diagonal length of the 25-m-resolution grid cells. In order to estimate stream water temperature throughout the network, we determined the relationship between annual mean water temperature (measured at the pCO_2 sites) and air temperature (WorldClim) using linear regression. We used this model (95 % CI in brackets, slope = 0.55 [0.53, 0.57], intercept = 5.08 [4.97, 5.18], R^2 = 0.59) to estimate stream water temperature, which in turn enabled us to calculate the gas transfer velocity (k) from k_{600} by the ratio of Schmidt numbers (Jähne et al., 1987). The F_{CO2}, reported as g C m^{-2} d^{-1}, was calculated using eq. 1 for the entire network with a pCO_{2sat} of 400 μatm. Stream segments intersecting inland water bodies in the EU-HYDRO data product were removed.

All data analysis was performed in R version 3.5 (R Core Team, 2018) and Python version 3.7 (Van Rossum & Drake, 2011). All code used in the analysis, the predictive model and resulting high-resolution grids (25 m resolution) with predicted stream pCO_2, *k* and F_{CO2} throughout the stream network have been deposited in an open repository.

3 Results

3.1 Stream pCO_2

Water chemistry data were collected across three Scandinavian countries, spanning a large geographical region from 54.5–70 °N and 8–31.5 °E (Fig. 1). The region covers a broad, north-temperate climate zone with mean annual air temperature ranging from -7–9.4 °C and very different degrees of anthropogenic influence and land use. Observations were spread evenly across the countries except for the northern parts of Sweden and Finland, where alkalinity were very close to or zero. We calculated annual mean pCO_2 and determined catchment characteristics for

2298 sampling stations. The average observed pCO_2 was 1494 (SD: 1426 and range: 17–15646) μatm; 13 % of the sites were below atmospheric saturation. The degree of supersaturation (mean pCO_2) was generally much more pronounced at sites in Denmark (3158, SD: 2403 and range: 438–15646 μatm) compared with Finland (1476, SD: 1035 and range: 17–8516 μatm) and Sweden (1069, SD: 960 and range: 91–9269 μatm; Fig. 1c). Annual mean alkalinity and pH were mostly higher in Denmark (2.5, SD: 1.3 and range: 0.12–6.8 meq L^{-1} and pH 7.6, SD: 0.4 and range: 6.1–8.4) relative to Sweden (0.7, SD: 1.0 and range: 0.12–6.0 meq L^{-1} and pH 7.2, SD: 0.3 and range: 6.1–8.3) and Finland (0.4, SD: 0.3 and range: 0.12–2.9 meq L^{-1} and pH 6.7, SD: 0.3and range: 6.1–9.4). High pCO_2 levels were found to occur in both high- and low-alkaline areas.

Martinsen, K. T., Kragh, T., & Sand-Jensen, K. (2020). Carbon dioxide partial pressure and emission throughout the Scandinavian stream network. Global Biogeochemical Cycles, 34, e2020GB006703. https://doi.org/10.1029/2020GB006703

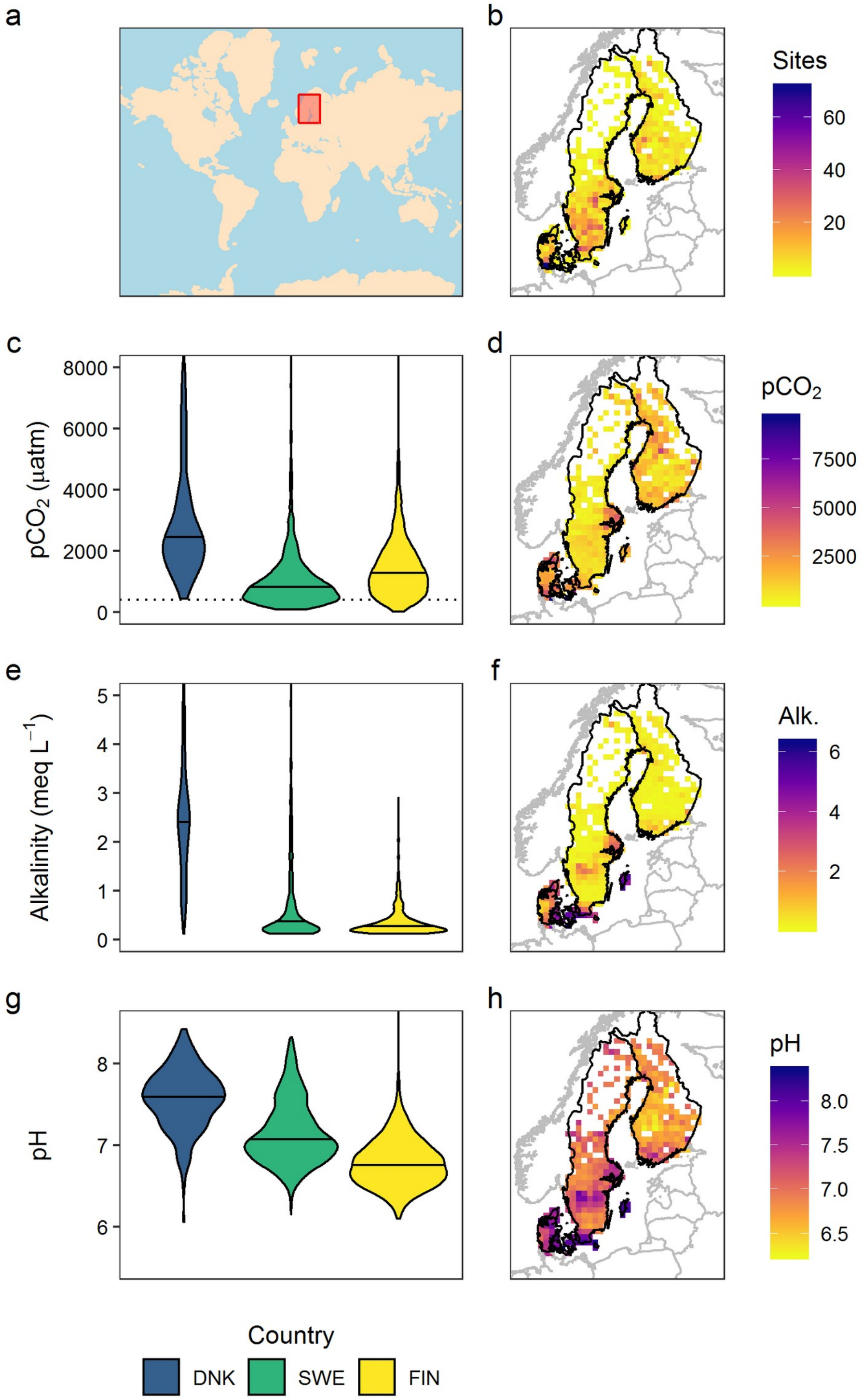

Figure 1. Density distributions and spatial variation of pCO_2, alkalinity and pH. a) World map showing the study region in red. Violin plots show the density distribution and median (horizontal line) of pCO_2 (c), alkalinity (e) and pH (g) by country. Gridded maps (40 km resolution) show the number of sites (b) and mean pCO_2 (d), alkalinity (f) and pH (h).

3.2 Predictive model selection

In order to predict pCO_2 throughout the stream network, we used nested cross-validation to compare the predictive performance of eight machine-learning models (Fig. 2). The simplest models (linear model, decision tree and k-Nearest Neighbor) did not perform well; the best of these was k-Nearest Neighbor. The Random Forest model showed the best performance, achieving the lowest RMSE and highest R^2. Generally, models with higher flexibility performed the best, especially tree-based ensemble models utilizing bagging or boosting (Random Forest and extreme gradient boosting). The distributions of predictive performance estimates showed some skewness but generally within a narrow range, suggesting decent model stability. The superior performance of the most flexible models indicates that the relationship between pCO_2 and catchment variables is complex, likely due to non-linearity and interactions between variables. In short, the more flexible models generated markedly more accurate predictions of pCO_2 than the traditional linear model.

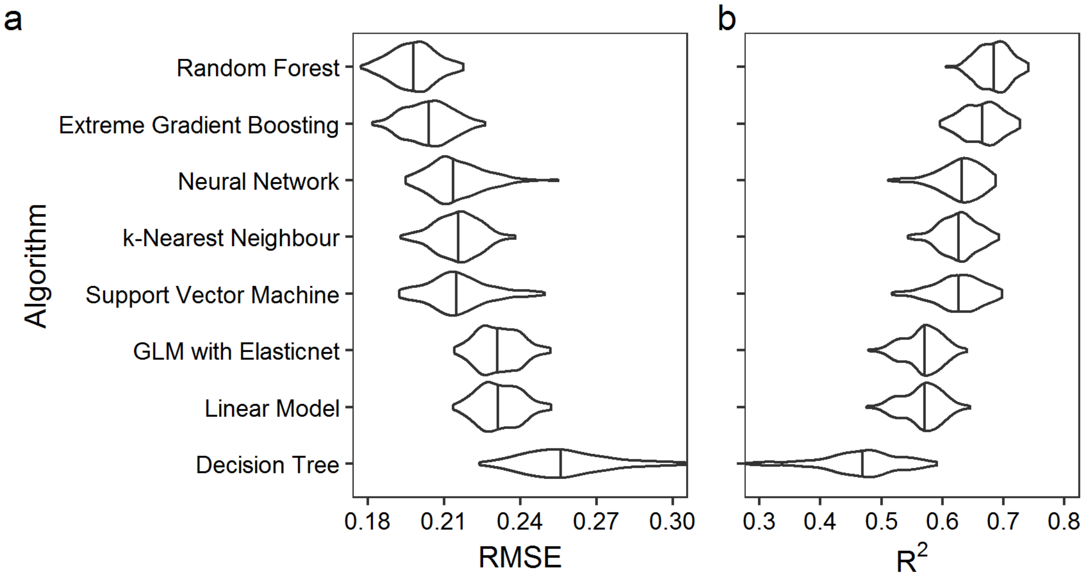

Figure 2. Violin plots showing the density distribution of 125 evaluations of predictive performance (5-fold cross-validation repeated 25 times) of eight candidate models as the root-mean-square error (RMSE; a) and explained variation (R^2; b). Models are sorted by decreasing predictive performance.

3.3 Predictive model performance

The Random Forest model trained and tuned on the training set, showed good performance on the test set with low RMSE (0.22 $\log_{10}(\mu atm)$) and high R^2-values (0.66; Fig. 3). These performance estimates are slightly worse than the estimates found during model selection but within the expected distribution (Fig. 2). The pCO_2 predictions clustered around the 1:1 line, with slight underestimation where pCO_2 was observed to be high and overestimation where it was observed to be low (Fig. 3).

Martinsen, K. T., Kragh, T., & Sand-Jensen, K. (2020). Carbon dioxide partial pressure and emission throughout the Scandinavian stream network. Global Biogeochemical Cycles, 34, e2020GB006703. https://doi.org/10.1029/2020GB006703

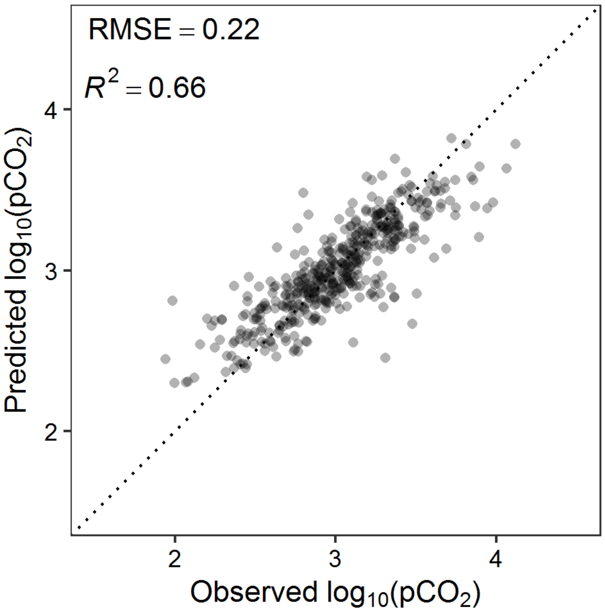

Figure 3. Performance of the final Random Forest model on the test set sample; shown as predicted (y-axis) versus observed (x-axis) values ($\log_{10}$ pCO_2). The dotted line is a 1:1 relationship.

Variables of most categories (except stream proximity) were included in the top six most important predictor variables in the final model (Fig. 4a): geo-morphometry (elevation, slope), land cover (permanent, temporary water cover and grassland) and climate (precipitation). The single most important variable was the average catchment elevation. The partial dependence plots show the response profile of the important variables in relation to pCO_2 (Fig. 4b). The response profiles of the most influential variables were non-linear. pCO_2 decreased with increasing catchment elevation and slope showing that the highest pCO_2 values are generally found in streams in lowland areas with flat terrain. Higher catchment precipitation, temporary water cover and especially the proportion of permanent water cover also resulted in lower stream pCO_2 levels. The catchment proportion of grassland had a positive influence on pCO_2 at low values.

The performance of the Random Forest model was worse when extrapolating beyond the geographical region used for training the model. Leaving one country out for testing while training the model on the two remaining resulted in much higher RMSE ($\log_{10}$(µatm), Denmark: 0.28, Sweden: 0.31 and Finland: 0.28) and poor R^2-values (Denmark: -0.09, Sweden: 0.09 and Finland: 0.06). A negative R^2 implies that model performance is worse than using the mean of the response.

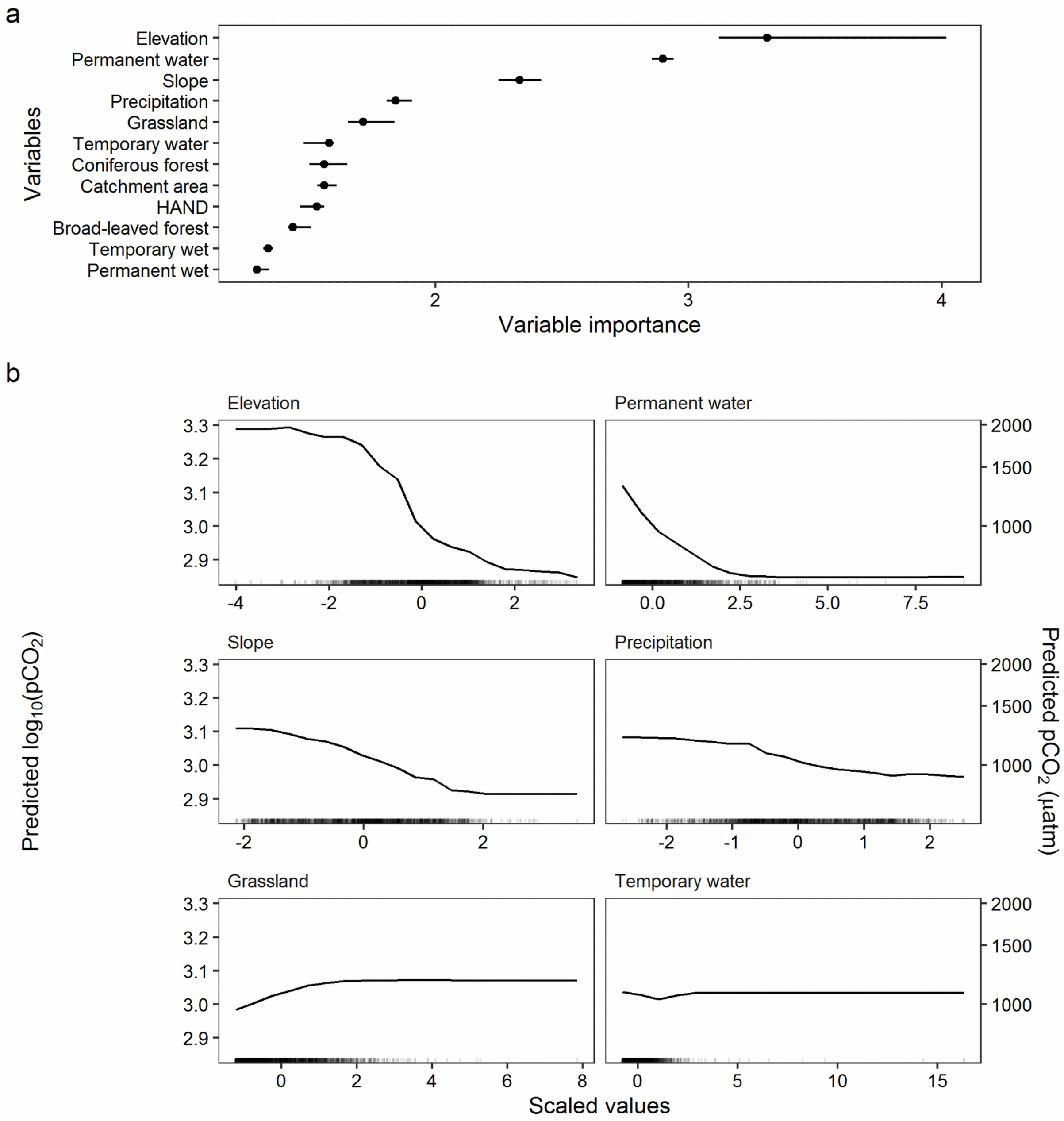

Figure 4. Variable importance (a) and partial-dependence plots (b) of the six most important variables in the Random Forest model. In b, the x-axis is the predictor variables after they have been transformed and scaled to unit standard deviation.

3.4 Estimated stream pCO_2 and F_{CO2}

We applied the trained Random Forest model to predict stream pCO_2 throughout the modeled stream network consisting of 7.8 million grid cells at a resolution of 25 m. For 7.6 million grid cells in this network we were also able to estimate F_{CO2} and *k* using supplementing data on runoff and air temperature. This modeled stream network visually appeared to correspond well with the network observed from very-high-resolution map imagery (Fig. 5a) and had an approximate length of 268.807 km (Denmark: 20.828 km, Sweden: 179.030 km and Finland: 68.949 km). The high-resolution pCO_2 network enabled local comparisons and yielded expected pCO_2 in a wide range of stream branches and confluences (Fig. 5b).

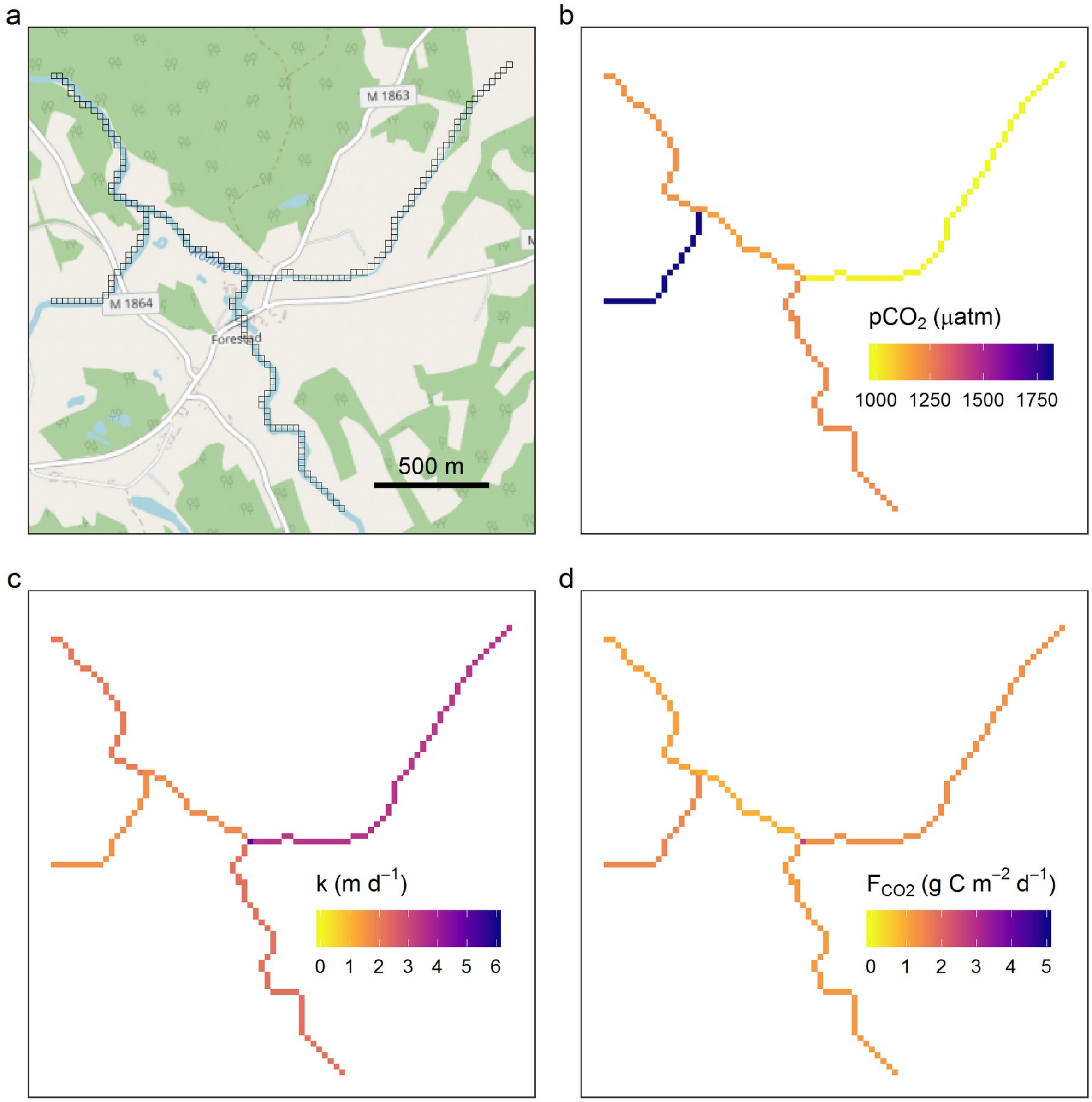

Figure 5. Example of the grids resulting from the analyses. Stream network on a map (OSM, 2020) showing the agreement between real stream lines and the network modeled and used here (a). Predicted pCO_2 (b), *k* (c) and F_{CO2} (d). The stream network resolution is 25 m and the depicted region is 4 by 4 km.

The mean of all pCO_2 predictions was 1134 (SD: 780 and range: 154–8174) μatm. The mean estimated *k* was 3.4 (SD: 4.1) m d^{-1} and F_{CO2} was 1.2 (SD: 1.4) g C m^{-2} d^{-1}. We calculated density distributions from 10 thousand random cells in the final grids (Fig. 6). The predicted pCO_2 was similar to the observations used for modeling (Fig. 1c), with a higher mean in Denmark (2721, SD: 900 and range: 592–8174 μatm) compared with Finland (1274, SD: 560 and range: 246–5640 μatm) and Sweden (896, SD: 584 and range: 154–5871 μatm; Fig. 6a). Mean estimated *k* values were generally low in Denmark (2.2 and SD: 1.4 m^{-1}) and Finland (2.0 and SD: 2.6 m^{-1}), while higher values were more common in Sweden (4.0 and SD: 4.6 m^{-1}; Fig. 6b). These opposites (high pCO_2, low *k* and vice versa) yielded more similar F_{CO2} distributions (Fig. 6C), though with higher mean emission rates in Denmark (3.2 and SD: 1.8 g C m^{-2} d^{-1}) compared with Finland (1.2 and SD: 1.4 g C m^{-2} d^{-1}) and Sweden (1.0 and SD: 1.1 g C m^{-2} d^{-1}).

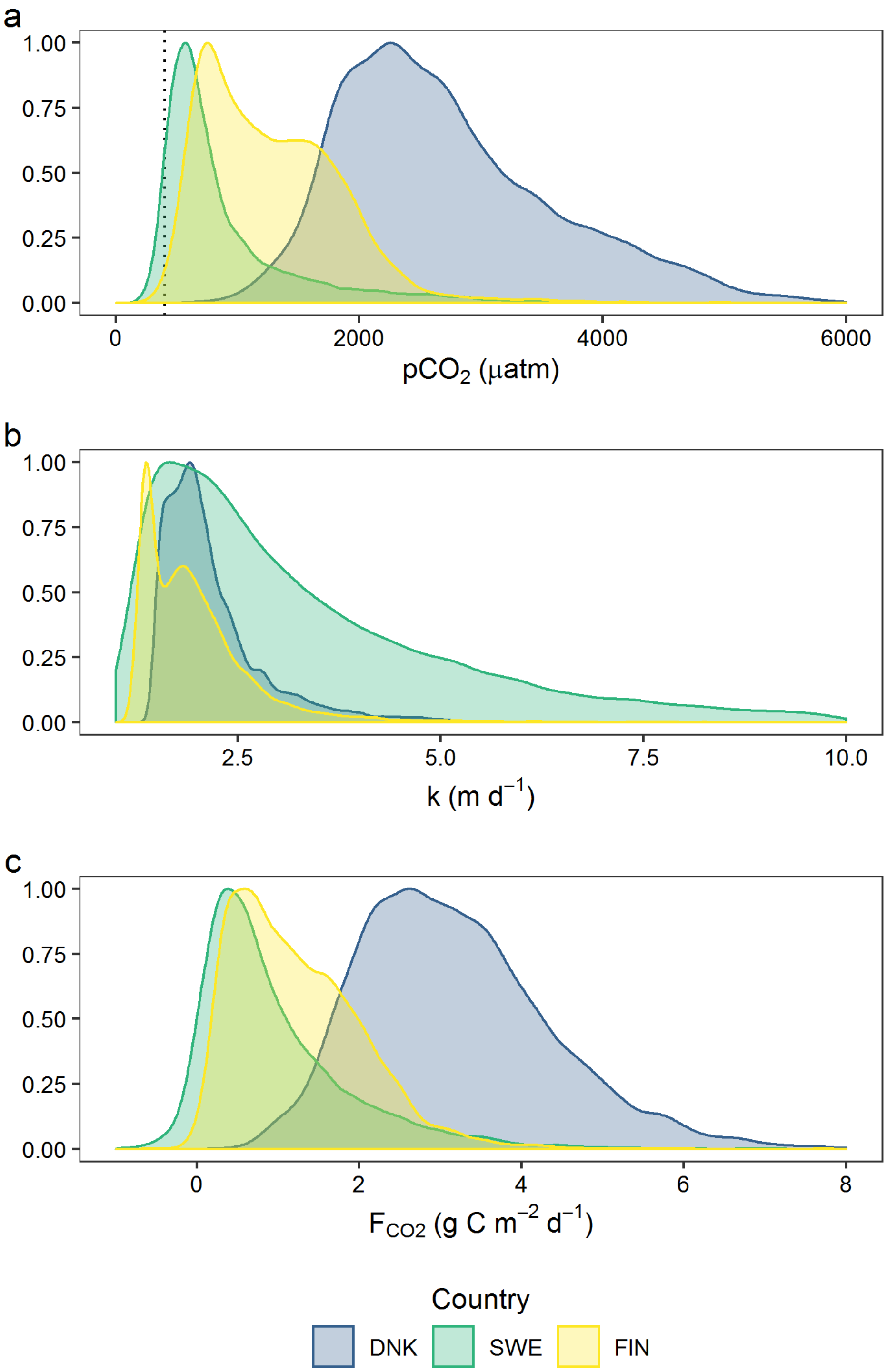

Figure 6. Density distributions (y-axis scaled to 1) of predicted pCO_2 (a), *k* (b) and F_{CO2} (c) sampled (n=10,000) from the final grids and colored by country. The vertical dotted line in A shows the atmospheric pCO_2. Tails of the distribution have been cut (maximum 2 % of observations) to improve visualization.

4 Discussion

4.1 Predicting pCO_2 in large stream networks

We combined open data from several sources and a Random Forest model to produce high-resolution maps of stream pCO_2 and estimates of *k* and F_{CO2} covering three Scandinavian countries. As hypothesized, stream pCO_2 can be predicted from catchment characteristics. Our use of a high spatial resolution of 25 m is an essential improvement, as it enables both local comparisons and calculations of more accurate national carbon emissions. Since the data required for our predictive model is obtained via remote sensing products whose use is widespread, our model can be used to predict pCO_2 levels over much larger geographical areas than is possible otherwise. Background data covering the European Union are already available and global coverage is likely within reach, which should make it possible to predict European and global pCO_2 levels and estimate F_{CO2} emissions from streams using our modeling approach. Expanding the predictive capacity to larger geographical regions will require more observations of pCO_2 covering a wider range of environmental conditions in order to train performant models. It is evident that leaving countries out during model training leads to poor predictive performance showing that extrapolation to new geographical regions is still difficult. It should be possible to achieve similar model performance when training new models with data for larger regions and it is not uncommon to see improvements in predictive performance when including more observations.

The primary goal of our analysis was to predict pCO_2 throughout a realistic model of a large stream network. Due to the scale chosen, we likely missed the smallest streams. A Danish stream inventory found that 19,260 km of streams were wider than 2.5 m (Sand-Jensen et al., 2006). Wallin et al. (2018) used a virtual stream network for Sweden in which streams of order 3 and above made up 95,353 km. Comparing these numbers with the stream lengths included in our analysis, lead us to believe that we missed only streams less than ~2.5 m wide, or of 1st and 2nd stream order. Furthermore, our stream lengths are likely conservative estimates due to bifurcation and sinuosity, which are approximated here as the diagonal of the grid cell size. Ultimately, our modeled stream network depends on data that is based on photo-interpretation of remote sensing imagery, where extraction of small streams is limited by both image resolution and obstructions such as forest canopies. These biases may provide an explanation for the discrepancy between the length of our network and that of Wallin et al. (2018). Modeling a realistic stream network that captures higher levels of detail than our models, will be possible with higher-resolution elevation models and finer-scale stream networks, which are likely available from national governmental agencies.

4.2 Models for predicting stream pCO_2

The mean of all our pCO_2 predictions (1134 μatm) is lower than Lauerwald et al.'s (2015) and Raymond et al.'s (2013) estimates of global stream pCO_2 of 2400 and 3100 μatm (indirect estimation of pCO_2), as well as those of Wallin et al. (2018), who estimated 2468 μatm based on direct measurements in Swedish low-order (1–4) streams. For Sweden, our mean resembles that of previous studies when streams of all sizes are included (Humborg et al., 2010; Weyhenmeyer et al., 2012), while being slightly lower when only small streams are included (Wallin et al., 2018). For this study region which is dominated by low-alkaline streams, the pH and organic alkalinity corrections of Liu et al. (2020) are important for reducing the bias of pCO_2 calculations and result in much lower values. Our model's predictions of pCO_2 achieved good accuracy (RMSE=0.22) and explained variation (R^2=0.66). This shows that, as we initially expected, a random forest model using only catchment predictors as input variables was a suitable model of stream pCO_2 in the network. This highlights the possibility of efficiently predicting annual CO_2 dynamics in large stream networks. The methodology might also be applicable for other carbon species such as methane or streams solutes in general which are believed to be driven by catchment scale processes.

Our approach to modeling pCO_2 and F_{CO2} in stream networks differs from previous studies primarily in the use of machine learning methods as opposed to linear models for predicting pCO_2. This contrasts with the simpler models and traditional linear models which have often been used in previous similar studies. For example, Lauerwald et al. (2015) trained a linear model with an R^2 of 0.47 using 1182 global samples. While the performance of this model is difficult to compare directly to the approach presented here, in part due to the different scales of the analyses, it is notable that part of the variation remains hard to explain, as also observed in other studies (Horgby et al., 2019; Humborg et al., 2010; Hutchins et al., 2020; Teodoru et al., 2009; Weyhenmeyer et al., 2012). A potential bias is introduced during modeling when a functional form of a relationship, e.g. linear, is assumed because as we also show, non-linear relationships between pCO_2 and predictors seem to be prevailing. Also, poor model fits could result from the omission of important predictors due to them being unknown or unavailable. Including drivers at multiple scales ranging from very proximate point processes, to catchments (as done here) and regions could likely improve models further. Hutchins et al. (2019) found that the inclusion of regional structure, correlated to terrestrial NPP and soil carbon, improved prediction of stream DOC, pCO_2 and methane. Including such regional baseline effects could be useful if regions or basins with distinctive carbon availability can be delineated. This should likely include subsurface processes that reflect pCO_2 in groundwater and geological influences such as weathering processes. Currently, it is difficult to identify the contribution on stream pCO_2 of such regional processes.

One option for improving the ability to predict pCO_2 in stream networks is to use more flexible models. However, flexibility comes at a cost and model tuning is necessary to avoid overfitting the training data, but the result is often more accurate predictions. We compared multiple machine learning models, and unsurprisingly the Random Forest model proved to be the best at generalizing the conditions of new samples. Many studies have found that this or similar types of models (extreme gradient boosting, gradient boosting machines, etc.) generally perform very well, also in settings with interactions, non-linearity and high dimensionality (James et al., 2013). The performance of linear models could likely be improved somewhat, compared to the simple way they were implemented here, by including interactions and quadratic terms. However, because structural relationships are complex and often unknown, it is left upon the researcher to uncover these relationships, which makes the modeling process difficult. Training a Random Forest model, we automatically approximate the functional relationships and attain high predictive accuracy (Breiman, 2001). As a downside, we are left to interpret the influence of predictor variables after model training. Since our end goal is improved predictability of stream pCO_2 on a large scale, sacrificing model interpretability for predictive performance seems like a reasonable choice.

4.3 Drivers of stream pCO_2

The responses of pCO_2 to the most important catchment variables are generally in agreement with previous studies. The influence of geo-morphometric variables such as catchment elevation and slope has been observed before (J. B. Jones & Mulholland, 1998b; Lauerwald et al., 2015; Smits et al., 2017), but the underlying mechanisms for these relationships remain unclear. Increasing catchment slope should intuitively decrease pCO_2 due to higher stream channel slope and, in turn, higher k and F_{CO2} in upstream reaches, but other mechanisms might also be relevant. Catchment slope may influence carbon loading in streams by acting as a proxy of wetland formation and soil accumulation (Smits et al., 2017). Lower slope and lower-altitude catchments are likely to have thicker soil profiles, be richer in organic matter and have lower hydraulic conductivity, resulting in higher water retention time in the catchment. This results in a higher quantity of organic matter, but of lower quality, which influences the in-stream processing of organic matter (Jankowski et al., 2014). Because catchment slope, and to some degree elevation, influences both k and carbon loading, this may explain its success as a predictor of stream pCO_2 on large scales. An increasing proportion of permanent water cover (primarily lakes) in the catchment also influences pCO_2 negatively. While the lake and stream pCO_2 may differ substantially, especially during summer (Weyhenmeyer et al., 2012), the influence of lake processes on downstream pCO_2 may diminish fairly quickly as their impact is overtaken by that of atmospheric exchange and new groundwater inputs (Crawford et al., 2014). However, it is likely that, due to long residence times accompanied by photosynthetic consumption, carbon retention and atmospheric emission of CO_2, pCO_2 is lower in streams whose catchment areas have higher proportions of lakes.

We derived a range of stream network proximity metrics, e.g., different stream orders and network lengths; as expected, all were highly correlated with catchment area characteristics, which was the only variable kept onwards for predictive modeling. We found that catchment area did not strongly influence pCO_2, however, it decreased with elevation, slope and permanent water suggesting that position along the stream network (e.g. downstream reaches) generally is important. This is not surprising, because as the stream-flow increases, so does the influence of in-stream metabolism relative to catchment contributions of CO_2 from soil respiration or groundwater inflow on stream pCO_2 (Hotchkiss et al., 2015). Furthermore, pCO_2 may decline downstream due to the ongoing emissions of CO_2 to the atmosphere, resulting in rapid declines on a scale ranging from meters to kilometers (Duvert et al., 2018). The influence of stream discharge, therefore, depends on the balance between its effect as a delivery mechanism of surface- and groundwater CO_2 — both directly and indirectly by organic matter (S. Liu & Raymond, 2018) — and as a sink that controls F_{CO2} through stream hydraulics (Long et al., 2015). Headwaters and smaller streams often have higher pCO_2 because they often are subsidized by terrestrial inputs; the influence of these inputs declines as stream size and discharge increase. Input of readily degradable organic matter does, however, contribute to CO_2 supersaturation, which was the prevailing state (87 % of sites) of the stream sites included in this analysis. One could expect oxygen and inorganic carbon to be in molar balance, but this is rarely the case (Torgersen & Branco, 2008). Instead, the imbalance could likely be used to examine the relative contribution of the catchment to in-stream processes (Vachon et al., 2020). The size-dependent metabolic scaling of streams appears similar to that of lakes, where the imbalance between the F_{oxygen} and F_{CO2} changes with lake surface area (Martinsen et al., 2019). A comparison of such integrating measures within and between ecosystems could provide new insight into the drivers of pCO_2 production and the “metabolic fingerprint” of streams.

4.4 Upscaling gas transfer velocity and flux

Using the predicted pCO_2 throughout the large stream network, we show how stream F_{CO2} can be calculated throughout the network by estimating *k* from empirical relationships. Such upscaling exercises are necessary to advance our understanding of streams in large scale carbon budgets. Increasing the temporal resolution from annual to monthly time scales might provide further improvements. Especially seasonal variations in hydrology and also ice cover, which we did not include in this analysis, are expected to have large influence on *k* and in turn F_{CO2}. High resolution data on runoff and remote sensing imagery could help alleviate these issues. The high influence of *k* on F_{CO2}, especially in steep and rugged terrain, makes it a reasonable target for new model refinements that could further improve estimation of large scale F_{CO2}, despite recent studies suggesting that the temporal variability is difficult to predict (Wallin et al., 2013).

Our work demonstrates the ease with which pCO_2 grids can be used to improve F_{CO2} estimates in stream networks across a large spatial scale. However, while the predictive pCO_2 model has been validated here, the uncertainties of the *k* and F_{CO2} estimates are unknown because validation is lacking. Interpretations of localized phenomena should thus be made more cautiously. Future studies should improve the estimation of these processes and focus on quantifying the associated uncertainty.

5 Conclusions

Using open and readily available data sources, we used a Random Forest model to produce a high-resolution grid of stream pCO_2. Using additional data on runoff and air temperature, we also estimated stream *k* and F_{CO2}. Due to its high resolution, our modeled stream network is both realistic and covers a large proportion of the region’s total stream network. We have shown how the use of more flexible predictive models can improve the accuracy of prediction. Furthermore, we have demonstrated how emissions of an important greenhouse gas can be predicted in a streams network using only catchment characteristics derived from remote sensing products. This suggests that our approach can easily be adapted for work on other regions and likely other carbon species. We have provided an approach to predict aquatic carbon species which could be relevant for future studies of carbon cycling and budgeting from catchment to global scale.

Acknowledgments

This study was supported by grants from COWIfonden to KSJ. We thank David Stuligross for linguistic corrections. We thank Marcus Wallin and Clément Duvert for their constructive comments which helped improve the manuscript. All scripts used in the analysis and the resulting grids are available from an open repository (link: https://sid.erda.dk/public/archives/2a69da119d14c8f29a7541a8fa901750/published-archive.html) and raw data are cited in the references.

References

Abril, G., Martinez, J.-M., Artigas, L. F., Moreira-Turcq, P., Benedetti, M. F., Vidal, L., et al. (2014). Amazon River carbon dioxide outgassing fuelled by wetlands. Nature, 505(7483), 395–398. https://doi.org/10.1038/nature12797

Abril, G., Bouillon, S., Darchambeau, F., Teodoru, C. R., Marwick, T. R., Tamooh, F., et al. (2015). Technical Note: Large overestimation of pCO2 calculated from pH and alkalinity in acidic, organic-rich freshwaters. Biogeosciences, 12(1), 67–78. https://doi.org/10.5194/bg-12-67-2015

Barbarossa, V., Huijbregts, M. A. J., Beusen, A. H. W., Beck, H. E., King, H., & Schipper, A. M. (2018). FLO1K, global maps of mean, maximum and minimum annual streamflow at 1 km resolution from 1960 through 2015. *Scientific Data*, *5*(1), 1–11. https://doi.org/10.1038/sdata.2018.52

Battin, T. J., Luyssaert, S., Kaplan, L. A., Aufdenkampe, A. K., Richter, A., & Tranvik, L. J. (2009). The boundless carbon cycle. *Nature Geoscience*, *2*(9), 598–600. https://doi.org/10.1038/ngeo618

Beck, H. E., de Roo, A., & van Dijk, A. I. J. M. (2015). Global maps of streamflow characteristics based on observations from several thousand catchments. *Journal of Hydrometeorology*, *16*(4), 1478–1501. https://doi.org/10.1175/JHM-D-14-0155.1

Berner, E. K., & Berner, R. A. (2012). *Global Environment: Water, Air, and Geochemical Cycles*. Princeton University Press, New Jersey.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., et al. (2016). mlr: Machine learning in R. *Journal of Machine Learning Research*, *17*(170), 1–5.

Bischl, B., Richter, J., Bossek, J., Horn, D., Thomas, J., & Lang, M. (2017). mlrMBO: A modular framework for model-based optimization of expensive black-box functions. *arXiv Preprint arXiv:1703.03373*.

Borges, A. V., Darchambeau, F., Lambert, T., Bouillon, S., Morana, C., Brouy‘ere, S., et al. (2018). Effects of agricultural land use on fluvial carbon dioxide, methane and nitrous oxide concentrations in a large European river, the Meuse (Belgium). *Science of the Total Environment*, *610-611*, 342–355. https://doi.org/10.1016/j.scitotenv.2017.08.047

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Butman, D., & Raymond, P. A. (2011). Significant efflux of carbon dioxide from streams and rivers in the United States. *Nature Geoscience*, *4*(12), 839–842. https://doi.org/10.1038/ngeo1294

Butman, D., Stackpoole, S., Stets, E., McDonald, C. P., Clow, D. W., & Striegl, R. G. (2015). Aquatic carbon cycling in the conterminous United States and implications for terrestrial carbon accounting. *Proceedings of the National Academy of Sciences*, *113*(1), 58–63. https://doi.org/10.1073/pnas.1512651112

Cole, J. J., Prairie, Y. T., Caraco, N. F., McDowell, W. H., Tranvik, L. J., Striegl, R. G., et al. (2007). Plumbing the global carbon cycle: Integrating inland waters into the terrestrial carbon budget. *Ecosystems*, *10*(1), 172–185. https://doi.org/10.1007/s10021-006-9013-8

Cory, R. M., Harrold, K. H., Neilson, B. T., & Kling, G. W. (2015). Controls on dissolved organic matter (DOM) degradation in a headwater stream: The influence of photochemical and hydrological conditions in determining light-limitation or substrate-limitation of photo-degradation. *Biogeosciences*, *12*(22), 6669–6685. https://doi.org/10.5194/bg-12-6669-2015

Crawford, J. T., Lottig, N. R., Stanley, E. H., Walker, J. F., Hanson, P. C., Finlay, J. C., & Striegl, R. G. (2014). CO_2 and CH_4 emissions from streams in a lake-rich landscape: Patterns, controls, and regional significance. *Global Biogeochemical Cycles*, *28*(3), 197–210. https://doi.org/10.1002/2013GB004661

Drake, T. W., Raymond, P. A., & Spencer, R. G. M. (2018). Terrestrial carbon inputs to inland waters: A current synthesis of estimates and uncertainty. *Limnology and Oceanography Letters*, *3*(3), 132–142. https://doi.org/10.1002/lol2.10055

Duvert, C., Butman, D. E., Marx, A., Ribolzi, O., & Hutley, L. B. (2018). CO_2 evasion along streams driven by groundwater inputs and geomorphic controls. *Nature Geoscience*, *11*(11), 813–818. https://doi.org/10.1038/s41561-018-0245-y

EEA. (2016). EU-DEM. Copernicus Land Monitoring Service. European Digital Elevation Model.

EEA. (2017). EU-Hydro. European Environment Agency. Copernicus Land Monitoring Service. European River Network Database.

Garousi-Nejad, I., Tarboton, D. G., Aboutalebi, M., & Torres-Rua, A. F. (2019). Terrain analysis enhancements to the height above nearest drainage flood inundation mapping method. *Water Resources Research*, *55*(10), 7983–8009. https://doi.org/10.1029/2019WR024837

Gattuso, J.-P., Epitalon, J.-M., Lavigne, H., & Orr, J. (2018). *seacarb: Seawater carbonate chemistry*.

GDAL, & OGR. (2018). *GDAL contributors. GDAL/OGR Geospatial Data Abstraction Software Library*. Open Source Geospatial Foundation.

Hijmans, R. J. (2019). *Raster: Geographic data analysis and modeling*.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, *25*(15), 1965–1978. https://doi.org/10.1002/joc.1276

Horgby, Å., Segatto, P. L., Bertuzzo, E., Lauerwald, R., Lehner, B., Ulseth, A. J., et al. (2019). Unexpected large evasion fluxes of carbon dioxide from turbulent streams draining the world’s mountains. Nature Communications, 10(1), 4888. https://doi.org/10.1038/s41467-019-12905-z

Hotchkiss, E. R., Hall Jr, R. O., Sponseller, R. A., Butman, D., Klaminder, J., Laudon, H., et al. (2015). Sources of and processes controlling CO_2 emissions change with the size of streams and rivers. *Nature Geoscience*, *8*(9), 696. https://doi.org/10.1038/ngeo2507

Humborg, C., Mörth, C.-M., Sundbom, M., Borg, H., Blenckner, T., Giesler, R., & Ittekkot, V. (2010). CO_2 supersaturation along the aquatic conduit in Swedish watersheds as constrained by terrestrial respiration, aquatic respiration and weathering. *Global Change Biology*, *16*(7), 1966–1978. https://doi.org/10.1111/j.1365-2486.2009.02092.x

Hutchins, R. H. S., Prairie, Y. T., & del Giorgio, P. A. (2019). Large-scale landscape drivers of CO_2, CH_4, DOC, and DIC in boreal river networks. *Global Biogeochemical Cycles*, *33*(2), 125–142. https://doi.org/10.1029/2018GB006106

Hutchins, R. H. S., Tank, S. E., Olefeldt, D., Quinton, W. L., Spence, C., Dion, N., et al. (2020). Fluvial CO2 and CH4 patterns across wildfire-disturbed ecozones of subarctic Canada: Current status and implications for future change. Global Change Biology, 26(4), 2304–2319. https://doi.org/10.1111/gcb.14960

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Jankowski, K., Schindler, D. E., & Lisi, P. J. (2014). Temperature sensitivity of community respiration rates in streams is associated with watershed geomorphic features. *Ecology*, *95*(10), 2707–2714. https://doi.org/10.1890/14-0608.1

Jähne, B., Münnich, K. O., Bösinger, R., Dutzi, A., Huber, W., & Libner, P. (1987). On the parameters influencing air-water gas exchange. *Journal of Geophysical Research*, *92*(C2), 1937–1949. https://doi.org/10.1029/jc092ic02p01937

Johnson, M. S., Lehmann, J., Riha, S. J., Krusche, A. V., Richey, J. E., Ometto, J. P. H. B., & Couto, E. G. (2008). CO_2 efflux from Amazonian headwater streams represents a significant fate for deep soil respiration. *Geophysical Research Letters*, *35*(17). https://doi.org/10.1029/2008GL034619

Jones, J. B., & Mulholland, P. J. (1998a). Carbon dioxide variation in a hardwood forest stream: An integrative measure of whole catchment soil respiration. *Ecosystems*, *1*(2), 183–196. https://doi.org/10.1007/s100219900014

Jones, J. B., & Mulholland, P. J. (1998b). Influence of drainage basin topography and elevation on carbon dioxide and methane supersaturation of stream water. *Biogeochemistry*, *40*(1), 57–72. https://doi.org/10.1023/A:1005914121280

Lauerwald, R., Laruelle, G. G., Hartmann, J., Ciais, P., & Regnier, P. A. G. (2015). Spatial patterns in CO_2 evasion from the global river network. *Global Biogeochemical Cycles*, *29*(5), 534–554. https://doi.org/10.1002/2014gb004941

Liu, S., & Raymond, P. A. (2018). Hydrologic controls on pCO_2 and CO_2 efflux in US streams and rivers. *Limnology and Oceanography Letters*, *3*(6), 428–435. https://doi.org/10.1002/lol2.10095

Liu, S., Butman, D. E., & Raymond, P. A. (2020). Evaluating CO_2 calculation error from organic alkalinity and pH measurement error in low ionic strength freshwaters. *Limnology and Oceanography: Methods*, 18(10), 606–622. https://doi.org/10.1002/lom3.10388

Liu, Y. Y., Maidment, D. R., Tarboton, D. G., Zheng, X., & Wang, S. (2018). A CyberGIS Integration and Computation Framework for High-Resolution Continental-Scale Flood

Inundation Mapping. *Journal of the American Water Resources Association*, *54*(4), 770–784. https://doi.org/10.1111/1752-1688.12660

Long, H., Vihermaa, L., Waldron, S., Hoey, T., Quemin, S., & Newton, J. (2015). Hydraulics are a first-order control on CO_2 efflux from fluvial systems. *Journal of Geophysical Research: Biogeosciences*, *120*(10), 1912–1922. https://doi.org/10.1002/2015JG002955

Marcé, R., Obrador, B., Gómez-Gener, L., Catalán, N., Koschorreck, M., Arce, M. I., et al. (2019). Emissions from dry inland waters are a blind spot in the global carbon cycle. *Earth-Science Reviews*, *188*, 240–248. https://doi.org/10.1016/j.earscirev.2018.11.012

Martinsen, K. T., Kragh, T., & Sand-Jensen, K. (2019). Carbon dioxide efflux and ecosystem metabolism of small forest lakes. *Aquatic Sciences*, *82*(1), 9. https://doi.org/10.1007/s00027-019-0682-8

Marx, A., Dusek, J., Jankovec, J., Sanda, M., Vogel, T., van Geldern, R., et al. (2017). A review of CO_2 and associated carbon dynamics in headwater streams: A global perspective. *Reviews of Geophysics*, *55*(2), 560–585. https://doi.org/10.1002/2016rg000547

McInerney, D., & Kempeneers, P. (2015). *Open Source Geospatial Tools: Applications in Earth Observation*. Springer International Publishing. https://doi.org/10.1007/978-3-319-01824-9

MFVM, & DCE. (2019). Surface water database. Ministry of Environment and Food of Denmark and Danish Centre for Environment and Energy.

MVM. (2019). Miljödata. Swedish University of Agricultural Sciences (SLU). National data host lakes and watercourses, and national data host agricultural land.

Olden, J. D., Lawler, J. J., & Poff, N. L. (2008). Machine learning methods without tears: A primer for ecologists. *The Quarterly Review of Biology*, *83*(2), 171–193. https://doi.org/10.1086/587826

OSM. (2020). OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org.

Pebesma, E. (2018). Simple features for R: Standardized support for spatial vector data. *The R Journal*, *10*(1), 439–446. https://doi.org/10.32614/RJ-2018-009

Raymond, P. A., Zappa, C. J., Butman, D., Bott, T. L., Potter, J., Mulholland, P., et al. (2012). Scaling the gas transfer velocity and hydraulic geometry in streams and small rivers. *Limnology and Oceanography: Fluids and Environments*, *2*(1), 41–53. https://doi.org/10.1215/21573689-1597669

Raymond, P. A., Hartmann, J., Lauerwald, R., Sobek, S., McDonald, C., Hoover, M., et al. (2013). Global carbon dioxide emissions from inland waters. *Nature*, *503*(7476), 355–359. https://doi.org/10.1038/nature12760

Rocher-Ros, G., Sponseller, R. A., Lidberg, W., Mörth, C.-M., & Giesler, R. (2019). Landscape process domains drive patterns of CO2 evasion from river networks. Limnology and Oceanography Letters, 4(4), 87–95. https://doi.org/10.1002/lol2.10108

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Sand-Jensen, K., & Staehr, P. A. (2012). CO_2 dynamics along Danish lowland streams: Water-air gradients, piston velocities and evasion rates. *Biogeochemistry*, *111*(1-3), 615–628. https://doi.org/10.1007/s10533-011-9696-6

Sand-Jensen, K., Friberg, N., & Murphy, J. (Eds.). (2006). *Running Waters: Historical development and restoration of lowland Danish streams*. Aarhus Universitetsforlag, Denmark.

Sand-Jensen, K., Pedersen, N. L., & Søndergaard, M. (2007). Bacterial metabolism in small temperate streams under contemporary and future climates. *Freshwater Biology*, *52*(12), 2340–2353. https://doi.org/10.1111/j.1365-2427.2007.01852.x

Smits, A. P., Schindler, D. E., Holtgrieve, G. W., Jankowski, K. J., & French, D. W. (2017). Watershed geomorphology interacts with precipitation to influence the magnitude and source of CO_2 emissions from Alaskan streams. *Journal of Geophysical Research: Biogeosciences*, *122*(8), 1903–1921. https://doi.org/10.1002/2017jg003792

SYKE. (2019). Open web services. Finnish Environment Institute SYKE.

Tarboton, D. G. (2017). Terrain analysis using digital elevation models (TauDEM) Utah Water Research Laboratory, Utah State University.

Teodoru, C. R., del Giorgio, P. A., Prairie, Y. T., & Camire, M. (2009). Patterns in pCO_2 in boreal streams and rivers of northern Quebec, Canada. *Global Biogeochemical Cycles*, *23*(2). https://doi.org/10.1029/2008gb003404

Torgersen, T., & Branco, B. (2008). Carbon and oxygen fluxes from a small pond to the atmosphere: Temporal variability and the CO_2/O_2 imbalance. *Water Resources Research*, *44*(2). https://doi.org/10.1029/2006WR005634

Vachon, D., Sadro, S., Bogard, M. J., Lapierre, J.-F., Baulch, H. M., Rusak, J. A., et al. (2020). Paired O_2-CO_2 measurements provide emergent insights into aquatic ecosystem function. *Limnology and Oceanography Letters*. https://doi.org/10.1002/lol2.10135

Van Rossum, G., & Drake, F. L. (2011). *The Python Language Reference Manual*. Network Theory Ltd., Massachusetts.

Wallin, M. B., Öquist, M. G., Buffam, I., Billett, M. F., Nisell, J., & Bishop, K. H. (2011). Spatiotemporal variability of the gas transfer coefficient (KCO_2) in boreal streams: Implications for large scale estimates of CO_2 evasion. *Global Biogeochemical Cycles*, *25*(3). https://doi.org/10.1029/2010gb003975

Wallin, M. B., Grabs, T., Buffam, I., Laudon, H., Ågren, A., Öquist, M. G., & Bishop, K. (2013). Evasion of CO_2 from streams - The dominant component of the carbon export through the aquatic conduit in a boreal landscape. *Global Change Biology*, *19*(3), 785–797. https://doi.org/10.1111/gcb.12083

Wallin, M. B., Campeau, A., Audet, J., Bastviken, D., Bishop, K., Kokic, J., et al. (2018). Carbon dioxide and methane emissions of Swedish low-order streams - A national estimate and lessons learnt from more than a decade of observations. *Limnology and Oceanography Letters, 3(3),* 156–167. https://doi.org/10.1002/lol2.10061

Weyhenmeyer, G. A., Kortelainen, P., Sobek, S., Müller, R., & Rantakari, M. (2012). Carbon dioxide in boreal surface waters: A comparison of lakes and streams. *Ecosystems*, *15*(8), 1295–1307. *https://doi.org/10.1007/s10021-012-9585-4*

Yeo, I.-K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*(4), 954–959. *https://doi.org/10.1093/biomet/87.4.954*

Zappa, C. J., McGillis, W. R., Raymond, P. A., Edson, J. B., Hintsa, E. J., Zemmelink, H. J., et al. (2007). Environmental turbulent mixing controls on air-water gas exchange in marine and aquatic systems. *Geophysical Research Letters*, *34*(10). *https://doi.org/10.1029/2006gl028790*