

# Geophysical Research Letters®



## RESEARCH LETTER

10.1029/2023GL107350

## Generalizing Tree–Level Sap Flow Across the European Continent

### Key Points:

- Long Short-Term Memory networks (LSTMs) can predict hourly sap flow for diverse trees and climates across Europe bridging tree level measurements to continental scales
- Training on the extensive SAPFLUXNET data set enhances the ability of the LSTMs to simulate sap flow accurately in time and space
- The study underscores the value of large-scale deep learning models and advocates for the expansion of data sets like SAPFLUXNET

### Correspondence to:

R. Loritz,  
Ralf.Loritz@kit.edu

### Citation:

Loritz, R., Wu, C. H., Klotz, D., Gauch, M., Kratzert, F., & Bassiouni, M. (2024). Generalizing tree–level sap flow across the European continent. *Geophysical Research Letters*, *51*, e2023GL107350. <https://doi.org/10.1029/2023GL107350>

Received 16 NOV 2023

Accepted 1 MAR 2024

### Author Contributions:

**Conceptualization:** Ralf Loritz, Chen Huan Wu, Maoya Bassiouni

**Investigation:** Ralf Loritz

**Methodology:** Ralf Loritz, Chen Huan Wu, Daniel Klotz, Martin Gauch, Frederik Kratzert, Maoya Bassiouni

**Software:** Chen Huan Wu

**Supervision:** Ralf Loritz

**Validation:** Ralf Loritz, Chen Huan Wu, Daniel Klotz, Maoya Bassiouni

**Visualization:** Ralf Loritz, Chen

Huan Wu, Martin Gauch, Frederik Kratzert

**Writing – original draft:** Ralf Loritz, Chen Huan Wu, Daniel Klotz, Maoya Bassiouni

**Writing – review & editing:** Ralf Loritz, Martin Gauch, Frederik Kratzert, Maoya Bassiouni

Ralf Loritz<sup>1</sup> , Chen Huan Wu<sup>1</sup> , Daniel Klotz<sup>2</sup> , Martin Gauch<sup>3</sup>, Frederik Kratzert<sup>4</sup>, and Maoya Bassiouni<sup>5</sup> 

<sup>1</sup>Institute of Water and Environment, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, <sup>2</sup>Helmholtz Centre for Environmental Research – UFZ, Department Computational of Hydrosystems, Leipzig, Germany, <sup>3</sup>Google Research, Zurich, Switzerland, <sup>4</sup>Google Research, Vienna, Austria, <sup>5</sup>Department of Environmental Science, Policy and Management, University of California, Berkeley, Berkeley, CA, USA

**Abstract** Sap flow offers key insights about transpiration dynamics and forest-climate interactions.

Accurately simulating sap flow remains challenging due to measurement uncertainties and interactions between global and local environmental controls. Addressing these complexities, this study leveraged Long Short-Term Memory networks (LSTMs) with SAPFLUXNET to predict hourly tree-level sap flow across Europe. We built models with diverse training sets to assess performance under previously unseen conditions. The average Kling-Gupta Efficiency was 0.77 for models trained on 50% of time series across all forest stands, and 0.52 for models trained on 50% of the forest stands. Continental models not only matched but surpassed the performance of specialized and baselines for all genera and forest types, showcasing the capacity of LSTMs to effectively generalize across tree genera, climates, and forest ecosystems given minimal inputs. This study underscores the potential of LSTMs in generalizing state-dependent ecohydrological processes and bridging tree level measurements to continental scales.

**Plain Language Summary** Transpiration is the dominant flux of water from the land surface to the atmosphere, yet it remains challenging to measure and estimate especially given different climates and tree types. This study shows how large-scale deep learning models can simulate sap flow, the movement of water within trees, with high precision, even in forests and conditions not previously studied. We show that these models excel when they learn from large and diverse data sets. The flexible design of our model training means that every new sap flow measurement can potentially be used to further optimize our networks. Our findings indicate that this approach of continuously updating the model with new information greatly improves its performance to simulate and predict tree-level sap flow. This work thus highlights the potential of deep learning models to generalize sap flow, thereby enhancing ecohydrological investigations from the tree to the continental scale.

## 1. Introduction

Accurate quantification of plant transpiration is a critical component in hydrological research, accounting for approximately 65% of global terrestrial evapotranspiration (e.g., Good et al., 2015). Plants play a pivotal role in controlling the exchange of water between the atmosphere and land surface. Yet, capturing complex plant water use responses to environmental conditions and estimating transpiration across spatio-temporal scale, remains challenging. Among the limited number of available measurements for plant transpiration, in-situ sap flow sensors are the most widespread technique due to their relative low cost and ease of use (Dugas et al., 1993). Sap flow has long been recognized, especially in ecology and plant physiology fields, as a fundamental measurement for deciphering vegetation functionality and transpiration dynamics in both forested (Granier & Loustau, 1994) and agricultural ecosystems (Dugas et al., 1994).

While the analysis of sap flow to gain process understanding or to improve predictions has been less prevalent in catchment hydrology than ecology, recent studies highlight how sap flow measurements can enhance understanding of intricate relationships among vegetation characteristics, hydrometeorological factors, and catchment properties. For instance, Hassler et al. (2018) conducted an extensive study to determine the relative influence of tree-, stand-, and site-specific characteristics on sap velocity patterns, using data from 61 beech and oak trees across 24 sites in Luxembourg. Their findings suggest that transpiration estimates at the catchment scale could be significantly improved by taking into account not just hydro-meteorological drivers, but also the spatial patterns

© 2024. The Authors.

This is an open access article under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

of the composition of forests (e.g., total basal area). Renner et al. (2016) showed that variability in sap flow driven by topography and aspect could be balanced out by the forest stand composition, resulting in equivalent transpiration rates across south and north facing hillslopes. This exemplifies how vegetation dynamics adapt to environmental conditions to effectively use available resources. Hoek van Dijke et al. (2019) used sap flow measurements to explore the link between normalized difference vegetation index (NDVI) and transpiration. They showed that NDVI is not always reliable for modeling transpiration, especially during drought periods, as the correlation between NDVI and sap flow can vary positively or negatively, influenced by seasonal changes, moisture availability, and hydrogeological factors. Integrating sap flow into catchment studies and understanding its spatio-temporal variability is thus key for improving transpiration estimates at landscape scales. There is therefore a critical need to generalize sap flow regionally beyond tree-level measurements.

Sap flow measurement campaigns have historically been designed to explore plant-soil interactions at plot or plant scales, as exemplified by studies such as Jackisch et al. (2020) and Seeger & Weiler, 2021. In an effort to amalgamate and harmonize these numerous, individual small-scale field studies into a comprehensive, global open-source sap flow database, the SAPFLUXNET initiative was established (Poyatos et al., 2021). The SAPFLUXNET database presents thereby a unique opportunity to learn generalizable relationships across different plant genera and different climates. This is essential for estimating sap flow in “ungauged” regions, forest stands, or specific trees where direct measurements of sap flow have not been conducted and plant water use is unmonitored.

Deep learning offers powerful avenues to find generalizations in large data sets (e.g., LeCun et al., 2015). For example, Koppa et al. (2022) employed a feed-forward neural network to analyze daily, global data sets from eddy covariance stations, satellites and sap flow measurements from SAPFLUXNET. Their aim was to create an accurate global vegetation stress model that improves simulations of the reduction of evaporation from its theoretical maximum, for instance during periods of water limitations. By integrating the feed-forward neural network into an existing process-based model, they improved the ability of the process-based model to estimate global evaporation rates. Further, Loritz et al. (2022) showed that recurrent neural networks could simulate vegetation dynamics in form of catchment-averaged sap velocities with low residuals even in areas where the model has not been trained. Similar to Koppa et al. (2022), Loritz et al. (2022) showed that the deep-learning-based sap velocity simulations can be coupled with a process-based, hydrological model, in this case with the objective to replace the semi-empirical Steward-Jarvis equation, resulting in more accurate transpiration estimates and ultimately better soil moisture simulations particularly during a drought year. In addition, Li et al. (2022), highlighted the ability of recurrent neural networks to estimate the vegetation dynamics of a specific tree species in New Zealand from standard meteorological variables. Both Loritz et al. (2022) and Li et al. (2022) found that particular recurrent neural networks like Long Short-Term Memory (LSTM; Hochreiter & Schmidhuber, 1997) models are suitable architectures to simulate and predict sap flow when they compared different deep learning architectures. However, both studies trained models on relatively small and local data sets, representing only the dynamics of a forest stand or a small catchment without deciphering different tree species behavior or different forest types. The extent to which LSTMs can detect consistent relationships between tree-level sap flow across different genera, measurement methods, climates and forests remains an open question.

Regionalized sap flow predictions are crucial information for understanding vegetation water use at larger scales, offering key insights into transpiration dynamics and plant-climate interactions. Such information is necessary to manage forest ecosystems, particularly under the evolving challenges of global change. Despite its importance, accurate sap flow data beyond the tree-level remains challenging due to measurement uncertainties and the complex interplay of various global and local environmental drivers and controls. This study aims to investigate the potential of using LSTMs for modeling tree-level sap flow at an hourly time scale across the European continent. We developed continental tree-level sap flow models leveraging the SAPFLUXNET database to extract generalized relations between sap flow from different tree genera, dynamic atmospheric drivers and forest stand characteristics. We developed different experimental training setups to evaluate the performance of these models in time and space. The presented deep learning approach offers avenues to overcome limitations of transpiration models when tree-level parameterizations for stomatal conductance are locally unavailable, could be used to assess different forest structures and their implications for regional transpiration rates, and ultimately provide robust sap flow based transpiration estimates across scales.

## 2. Material and Methods

### 2.1. Data—A European Subset of the SAPFLUXNET Database

The SAPFLUXNET database (Poyatos et al., 2021) represents a comprehensive global repository of tree-level sap flow measurements and their ancillary data including tree and forest characteristics and meteorological observations. We used version 0.1.5 of the database and selected a European subset comprising 64 forest stands out of 202 stands in the global data sets, encompassing 738 individual trees. We specifically focused on the European subset of SAPFLUXNET due to the high density of measurements and strong overlap of tree genera found within this region. In total, we included six tree genera with >20 individual tree measurements: 282 plants with *Pinus*, 159 plants with *Picea*, 144 plants with *Quercus*, 94 plants with *Fagus*, 30 plants with *Larix*, and 29 plants with *Pseudotsuga*. Selected data sets represented six forest types according to the International Geosphere-Biosphere Programme (IGBP) classification: 34 evergreen needle-leaf forest (ENF), 11 mixed forest (MF), 8 deciduous broadleaf forest (DBF), 5 deciduous needle-leaf forest (DNF), 4 evergreen broadleaf forest (EBF), and 2 savannas (SAV). We treated each individual tree's seasonal sap flow time series separately resulting in a total of 2,279 years of sap flow ( $\text{cm}^3 \text{h}^{-1}$ ) observations, each corresponding to about one growing season (April to September, 3–6 months). The division into individual seasonal time series is a key in this study as it allows us to include all sensors, even if they cover only a few months at any possible point in time. The winter period (October to March) was excluded because transpiration is low or zero at most stands.

#### 2.1.1. Selected Model Features

We considered six dynamic features comprising meteorological variables at an hourly resolution available in the SAPFLUXNET database: air temperature ( $^{\circ}\text{C}$ ), relative humidity (%), vapor pressure deficit (kPa), shortwave incoming radiation ( $\text{W m}^{-2}$ ), precipitation (mm), and wind speed ( $\text{m s}^{-1}$ ). We considered six static features (time invariant), comprising four forest stand characteristics: mean elevation (m), long-term mean annual temperature ( $^{\circ}\text{C}$ ), long-term mean annual precipitation (mm), and forest type (DBF, DNF, EBF, ENF, MF, SAV), and two individual tree-level characteristics: diameter at breast height (DBH; cm) and tree genera (*Fagus*, *Larix*, *Picea*, *Pinus*, *Pseudotsuga*, and *Quercus*). We implemented one-hot-encoding for each of the 6 genera and 6 forest types in the data set. Our feature selection was identified by several trial-and-error runs to balance model performance and practical applicability. This selection reflects our strategic choice to minimize input variables and use the maximum number of sites in Europe and to enhance the model's generalizability across unmeasured forest stands, where such detailed information might not be readily available. Omitted variables, which can be important sap flow predictors (e.g., soil water limitation, total basal area) may increase model performance if data becomes more prevalent across sites in the future.

### 2.2. Deep Learning Model—Long Short-Term Memory

LSTMs are recurrent neural networks that are specifically engineered to circumvent the vanishing and exploding gradient problem encountered in regular recurrent neural networks (Hochreiter, 1998; Hochreiter & Schmidhuber, 1997). This is achieved through the introduction of a cell state, which provides the network with the capability to learn long-term dependencies that are typically important in environmental, sequential data. The memory cell works in conjunction with so-called “gates,” mechanisms that evolve the memory and output over time, while allowing the error to propagate consistently through the network, thereby facilitating the learning process. LSTMs have showcased their efficiency and aptitude in hydrological modeling and have emerged as one of the top-performing models for simulating various state-dependant ecohydrological phenomena, such as streamflow, soil moisture and ecosystem water and carbon fluxes (e.g., Besnard et al., 2019; De Bartolomeis et al., 2023; Kratzert et al., 2019). We chose LSTMs in this study following the results of Loritz et al., 2022 who previously showed that LSTMs and Gated recurrent networks (GRUs; Cho et al., 2014) are suitable model architectures for simulating hourly sap velocity.

#### 2.2.1. LSTM Hyperparameters

We explored ranges of hyperparameter settings based on previous hydrological studies (Mai et al., 2022) and found that LSTM performance was relatively stable with a wide range of hyperparameters. We chose settings from Mai et al., 2022, with modifications made to (a) a reduced sequence length (which represents the number of time steps the network looks back to process and learn from sequential inputs) and (b) a reduced number of

epochs. Both changes only minimally influenced model performance while greatly decreasing training times. The hyperparameter setting used for all model variants are: number of hidden layers = 1; hidden layer neurons = 256; learning rate = 0.0005; dropout rate = 0.4; batch size = 64; sequence length = 24 hr; epoch 20; iterative optimization algorithm = ADAM.

### 2.2.2. Model Setups and Evaluation

We developed several experimental model setups (see Table 1) using different amounts of data for model training to test the performance of the LSTMs to predict hourly sap flow in time and space. Our objective is to compare these model setups (a) to quantify model performance in predicting sap flow for unseen periods and for unseen locations and (b) to assess the value of training LSTMs on larger data sets like SAPFLUXNET instead of building models individual for each forest stand or genera. In the following we explain each model setup in detail:

We developed *gauged-continental models*, trained across all 64 forest stands to assess the ability of the LSTMs to generalize in unseen time periods but in seen forest stands (temporal performance). We developed *ungauged-continental models*, trained on random subsets of 33 forest stands and tested on 31 stands, to examine the LSTM's ability to generalize and predict sap flow at unseen forest stands (spatial performance). We developed two types of specialized models (*single forest stand models*; *single tree genera models*) on smaller, more constrained training sets, to infer if the LSTM's performance improves if trained on a large and diverse database like SAPFLUXNET. *Single forest stand models* are trained only on a single forest stand (location) across each genera. *Single tree genera models* are trained on a single genera (e.g., Fagus, Pinus) but across all forest stands where the genera is present. Further, we developed *gauged* and *ungauged baselines* using a simple statistical approach to benchmark the LSTM performance. *Gauged baselines* represent the monthly averaged hourly diurnal cycle of sap flow for each stand and for each genera present at a stand across the European continent. *Ungauged baselines* represent the monthly averaged hourly diurnal cycle of sap flow for each genera across the European continent. These baselines are built 10 times on the same randomly selected training data as the *gauged-* and *ungauged-continental models* and are exclusively derived from the data without using a machine learning model.

For all model setups, data were divided such that a single vegetation growing season of tree-level sap flow data is treated as an individual time series and entirely part of either the training, validation or test data (in total 2,279 individual time series). For the *gauged-continental models* (Table 1) we split the 2,279 individual time series 10 times into 1,140 years (50%) for training, 912 years (40%) for testing and set aside the remaining 227 years (10%) for validation. The *ungauged-continental models* were developed using a training data set composed of 10 randomly selected subsets, each containing 33 forest stands of the totally available 64 forest stands. These models were then tested on the remaining separate testing set comprising 31 forest stands. To ensure representativeness, the division of training and testing for the *ungauged-continental models* was stratified, maintaining proportional representation of each forest type as classified by the IGBP. This approach guaranteed that all IGBP forest types were consistently included in each training subset.

By training and testing all models in this study on 10 randomly drawn subsets without repetitions using a so-called Monte Carlo cross validation scheme, we can assess the robustness of our models with respect to the information content of the training data (Maier et al., 2023). Further, it allows us to train network ensembles to assess uncertainties of our sap flow simulations. All numeric input features were standardized by subtracting the mean and by dividing them by the standard deviation of the training data. We assessed each model's performance using the Kling-Gupta efficiency (KGE, Gupta et al., 2009) and its three components: Pearson correlation ( $r$ ), bias ratio ( $\alpha$ ), and variability ratio ( $\beta$ ), the Nash-Sutcliffe efficiency (NSE), and the mean absolute error (MAE).

## 3. Results

### 3.1. Performance of Gauged-Continental Models

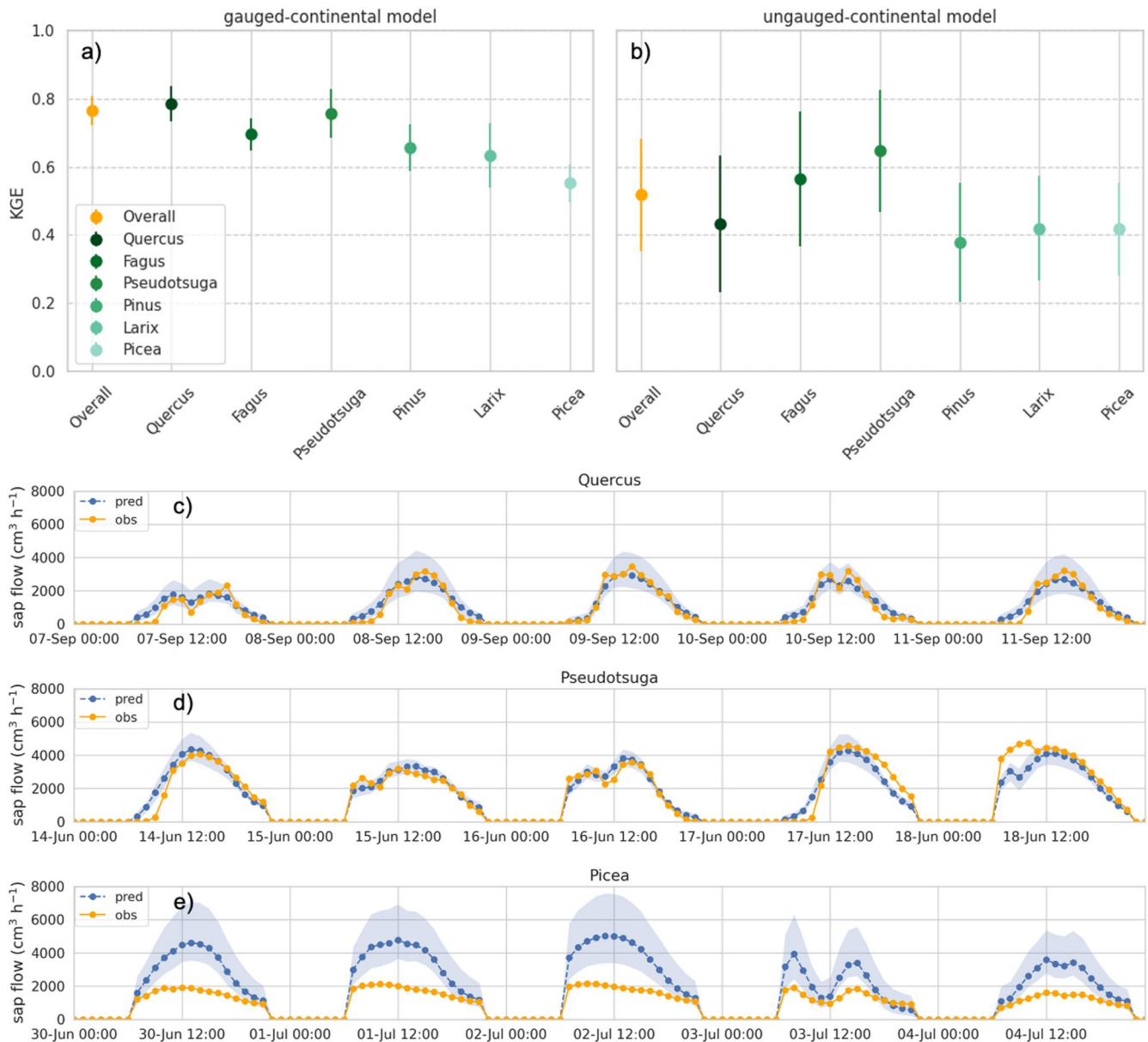
The *gauged-continental models*, representing the upper bound in continental performance, were capable of simulating tree level hourly sap flow with an average KGE of  $0.77 \pm 0.04$  (Figure 1a), in comparison to  $0.64 \pm 0.05$  for the *gauged baselines* (long-term monthly averaged hourly diurnal sap flow cycle for each stand for each genera, Table 1). Looking closely at model performance across tree genera, there is a consistent pattern of high KGE values for Quercus, Fagus, and Pseudotsuga trees. Notably, even Pseudotsuga, the tree genera with one of the smallest data set (80 years), achieves a KGE of  $0.76 \pm 0.07$  in contrast to a KGE of  $0.34 \pm 0.07$  for the

**Table 1**  
*Description of Model Training Setups From Specialized to Regionalized Approaches*

Model setup name	Description	Evaluation purpose
Single forest stand models	LSTMs trained using data exclusively from one forest stand, incorporating a variety of tree genera if available at the site	To evaluate the effectiveness of using continental data sets like SAPFLUXNET for model training versus training LSTMs on local data sets, for instance from one research project
Single tree genera models	Focused on a single tree genera (e.g., Fagus, Pinus) at a time, trained across multiple forest stands where this genera is found	To assess the benefit of training LSTMs on continental data sets that include several genera at the same time over models dedicated to only simulate a single genera
Gauged-continental model	LSTM trained on random subsets (50%) of the full data set covering all forest stands. Tested on the remaining unseen growing seasons	To assess performance of the LSTM for unseen time periods (time) at known forest stands, representing the upper bound in ungauged LSTM performance
Gauged baselines	Long-term monthly averaged hourly diurnal sap flow for each genera at each location. Estimated at each location with the same training data as the gauged-continental models	To assess performance for unseen time periods at known forest stands, offering a “purely” statistical benchmark for the performance of the gauged-continental LSTM
Ungauged-continental model	Trained on 33 forest stands and tested on the remaining 31, with random subsampling constrained by forest type (IGBP)	To determine model efficacy for entirely unseen forest stands (space)
Ungauged baselines	Long-term monthly averaged hourly diurnal cycle of sap flow for each genera across Europe. Built on the same training data as the ungauged continental models (33 forest stands)	To evaluate the performance for unseen time periods at known forest stands, serving as a benchmark for the performance of the ungauged-continental LSTM

*Note.* All models are trained, validated, and tested in an hourly resolution.





**Figure 1.** Performance and standard deviation ( $\pm$ ) of the 10 (a) gauged-continental models and (b) the ten ungauged-continental models measured by the KGE for all testing data (overall; orange) and for each of the six tree genera (genera name; green). Observed and simulated hourly sap flow using the gauged-continental model for five consecutive days in the testing data for (c) a Quercus tree in the year 2011 (France), (d) a Pseudotsuga tree in the year 2012 (Germany) and (e) for a Pinus tree in the year 2001 (Sweden). Blue bands visualize the uncertainty of the gauged-continental model given the Monte Carlo subsampling scheme and the individually trained LSTMs.

*gauged baselines*. Sap flow simulations of the Picea trees showed weaker performances with KGEs of  $0.55 \pm 0.06$ , despite being frequently found in the data (*gauged baselines* =  $0.28 \pm 0.03$ ). The amount of data for a tree genera does not directly correlate with the model performance.

Figures 1c, 1d, and 1e illustrate three sequences of hourly sap flow observations and simulations based on the *gauged-continental models* of a Quercus, Pseudotsuga and Picea tree for five consecutive days. While there is some uncertainty in the simulations (blue band), most models of the ensemble agree on the diurnal sap flow pattern, which is underpinned by the fact that Pearson correlations between different members of the model ensemble are all higher than 0.8. In agreement with the overall findings, the *gauged-continental models* capture the sap flow dynamics from the shown Quercus tree well (Pearson correlation = 0.9) and also match the absolute

values ( $\beta = 1.03$ ). The model performance for the shown Picea tree was lower and matches the dynamics to a certain extent (Pearson correlation = 0.77) but misses the absolute values. For the Pseudotsuga, the patterns and absolute values are well matched and even the drop of sap flow during midday for 2 days is matched reasonably well by the *gauged-continental models*. Model uncertainty across random subsampling is low for both the Quercus and Pseudotsuga tree compared to the Picea tree.

### 3.2. Performance of Specialized Models

*Single tree genera models* trained on a single genera across all sites (Table. 1) did not outperform the *gauged-continental models*. At best they achieved an equivalent performance. Further, *single tree genera models* were found to be more sensitive to the amount of data, particularly if trained only on a few sites, and can exhibit large performance differences, even leading to negative KGEs at some locations. In contrast, the *gauged-continental models* remain relatively stable ( $\pm 0.04$  KGE) and are less affected if the data representing a genera is reduced or if the number of stands is varied. Here, no simulation, not even removing a tree genera completely, resulted in a negative KGE. This indicates that (a) the randomly selected training data due to the Monte-Carlo subsampling of the *gauged-continental models* each hold a similar amount of information about the relationship between input features and sap flow and (b) it opens avenues to extend the training data set by new tree genera even if they have only been measured at a few locations for a short period.

We also compared the outcomes of the *single forest stand models* and *gauged-continental models*, focusing on the specific tree genera measured at each site as not all tree genera are present in each forest stand. At certain forest stands, for example, in some locations in central Europe, the performance of the *single forest stand models* closely mirrors the *gauged-continental models* for the dominant genera at these sites (Fagus). However, in some locations in South France, performance was lower than the *gauged-continental models* for the dominant genera (Quercus trees) with KGEs around 0.5 or 0.6 versus KGEs of 0.79. However, despite being trained on more consistent data, owing to the fact that the same team managed each site and primarily the same types of sensors were used for sap flow measurement, no *single forest stand model* exceeded the performance of the *gauged-continental models* in any of the forest stands.

### 3.3. Performance of Ungauged-Continental Models

As expected, the performance of the *ungauged-continental models* is on average lower than that of the *gauged-continental models* yet proved reasonable with an average KGE of  $0.52 \pm 0.16$  in comparison to a KGE of  $-0.11 \pm 0.15$  of the *ungauged baselines*. We found that the performance of the models was particularly affected by the frequency of forest types, despite stratifying random subsampling of the training data by IGBP. In other words, forest type with lower frequencies in the training data had a larger effect on the overall test scores than the amount of samples of a tree genera. The observed performance variance of the *ungauged-continental models* (standard deviation of 0.16) highlights the variability of the information content in the training data due to the random subsampling. This variability reduces if the model is trained on more observations and on a wider range of forests. For instance, model performance of the *ungauged-continental models* increased rather quickly while the standard deviation of the ensemble dropped if, for instance, 70% instead of 50% of the data was used for training the *ungauged-continental models* (KGE  $\sim 0.65 \pm 0.08$ ).

## 4. Discussion

### 4.1. Gauged-Continental Models: Robust Sap Flow Estimates Across Europe, Overcoming Measurement Uncertainties and Generalizing Across Climates and Tree Genera

In this study, gauged-continental models always matched or outperformed specialized single genera or stand models. Given that SAPFLUXNET uses various sensors and measurement techniques across the globe (see the discussion of the SAPFLUXNET publication during the review process in Earth System Science Data; Poyatos et al., 2021) it is striking that the *gauged-continental models* trained in this study can generalize tree-level sap flow across diverse climate zones and genera with a performance akin to the specialized models that have been trained and tested on more consistent data set in this study (e.g., in terms of sensor type, installing method, tree type, climate zone) and compared to the results described by Li et al. (2022), Loritz et al. (2022).

Gathering sap flow data can result in strong variations even at the same tree, making it challenging to achieve consistent and accurate readings (Steppe et al., 2010). For instance, sap flow measurements taken just a few meters apart in trees of similar genera, size, and height can show significant differences (Vandegehuchte & Steppe, 2013). Particularly absolute values vary, while the overall dynamics, akin to what is frequently found with respect to in-situ soil moisture observations, are typically well-matched if similar sensors are installed in different trees located in the same forest stand (e.g., Hassler et al., 2018; Loritz et al., 2017; Zehe et al., 2010). Differences in absolute values might thereby arise from various small-scale structural characteristics, such as properties of the sap wood or heterogeneous flow paths inside the stem. These factors are typically unknown and not provided to our models as input. The model has, therefore, no way to learn such small-scale differences that might explain why at two similar trees in close proximity different absolute sap flows have been measured. This might be one reason why all models in this study generally learn to represent the dynamics of sap flow well, but can exhibit a bias for certain trees.

Our findings reinforce previous research suggesting that deep learning models, trained on large and diverse data sets, often outperform those trained on more specialized, homogenous data (e.g., Kratzert et al., 2019; Wi & Steinschneider, 2022). The latter depends, however, on the information content of the database which tends to, but does not have to, increase with data set size (Singh & Bárdossy, 2012). The results of our study, as shown in Figures 1c, 1d, and 1e, demonstrate that the Monte Carlo subsampling method developed specifically for our research goal to simulate sap flow, along with the application of network ensembles, significantly enhances our ability to address uncertainties in our model simulations. This approach aligns with recent advancements in data analysis, such as the use of Gaussian mixture models as a headlayer for LSTMs, a technique effectively employed by Klotz et al. (2022) for analyzing rainfall-runoff data. A technique we plan to explore to assess uncertainties within the SAPFLUXNET data set in our future research.

#### 4.2. Ungauged-Continental LSTMs Provide Reasonable Sap Flow Estimates at Ungauged Forest Stands, and New Data Have the Potential to Further Enhance This Capability

We quantified LSTMs performance for predicting hourly sap flow at forest stands that were unseen during training. The results show that the *ungauged-continental model's* performance, although lower than that of the *gauged-continental model*, was nevertheless still reasonable with an average KGE of around 0.52. We found that the best-performing forest stands were often in the most frequent forest types (e.g., ENF). While the model was capable of predicting sap flow also in boreal and mediterranean forests, where measurements are more scarce, these predictions became less reliable the further they deviated from the training data showing clear limitations of the *ungauged-continental model* and the chosen training data set. The random subsampling of the training data and experiments with increasing the training data highlight that each new set of sap flow data can make the LSTM more robust, particularly in vegetation types that are less frequently monitored. Given that there are many tax funded large sap flow data sets available in Europe (and likely in other parts of the world) that are yet not openly available and not included in SAPFLUXNET, we argue that our study hints toward the currently unused potential that lies ahead when these data sets are shared in a consistent manner or if new measurement campaigns would be designed specifically to close the spatial and ecological gaps in the SAPFLUXNET data set. The methodology we employed to segment and aggregate sap flow data into distinct time series based on growing seasons ensures that all future sap flow measurements can be incorporated into our model, provided they exceed the minimum sequence length (>24 hr) required by the LSTM.

The *ungauged-continental models* can make reasonable hourly predictions of sap flow in unseen forest stands for a majority of the European continent. This entails that with prescribed forest type and weather data (e.g., from climate models), it is possible to create hourly sap flow maps for Europe with one, continental deep learning model. Surely these dynamic maps have clear limitations but there are limited options to gather hourly information about plant water use at the tree level at ungauged sites. Furthermore, as we simulate tree-level sap flow based on different forest stand characteristics (DBH, genera) this model could be used to assess different forest structures and their implications for regional transpiration rates and how they potentially change under different forest management strategies. Such dynamic sap flow maps could also be used to evaluate or replace transpiration models that are frequently found in land surface or hydrological models as shown in Loritz et al. (2022). Our study hence demonstrates an avenue toward developing ensembles of continental sap flow models to predict sap flow and evaluate vegetation dynamics around the globe. Future research could explore additional deep learning



architectures for sap flow prediction and refine feature selection in SAPFLUXNET, potentially integrating satellite data or global reanalysis products to estimate sap flow in forest without ground measurements.

### 4.3. Impact of Local and Global Environmental Factors on Model Simulations

The two continental models showed lower variance during random subsampling of training data than the specialized models. Furthermore, they could simulate sap flow for tree genera not included in the training set, achieving a KGE between 0.2 and 0.4. The latter suggests that forest type information (IGBP) is more crucial for the models than specific tree genera data. This could be due to lesser variability within certain forest types across different tree genera compared to the variability across forest types.

The SAPFLUXNET database includes sap flow measurements with various sensor types, which introduces different levels of uncertainty and affects comparability (e.g., Köstner et al., 1998; Lundblad et al., 2001; Tournebize & Boistard, 1998). However, incorporating sensor type information into our models did not reduce simulation errors nor did the performance for certain types of sensors improve compared to other sensors. This finding suggests that differences in sap flow measurements across sensor types are not systematic but rather somewhat random and underpins the findings of Tournebize and Boistard (1998) that sap flow taken with different measurement techniques in forests is, besides the inherent uncertainty of every sap flow measurement, comparable.

Our analysis emphasizes the significance of tree and forest site characteristics in explaining variations in sap flow, in agreement with the findings of Hassler et al. (2018). Specifically, during our feature selection, we observed that forest type (IGBP) notably improved model performance, while the addition of total stand density or total basal area slightly increased simulation accuracy (KGE increased from 0.79 to 0.81). Given the marginal improvement, we opted to exclude these features from the final model to ensure its wider applicability. Our study encompasses a broad range of climate zones and forest types and show, unlike catchment-level studies (e.g., Hassler et al. (2018)), that climatic factors, particularly long-term average precipitation and temperature, play a crucial role in our models. They are key to explaining variability in plant water uptake and water use strategies across different biomes (Bassiouni et al., 2023).

It is important to note that our findings are contingent on the current state of the SAPFLUXNET data set. As this data set expands, incorporating measurements from similar climate and forest types, factors like the hydro-pedological setting and more detailed forest stand descriptions may gain importance. Further, exploring the contribution of individual variables through advanced interpretability methods could shed light on their relative importance, serving also as a validation for the model. While this approach is compelling and merits further investigation, it requires a specialized training process for precise evaluation, which is beyond the scope of this study.

## 5. Conclusions

This study evaluated the potential and limits of deep learning to generalize sap flow dynamics across the European continent using the SAPFLUXNET database and effectively bridge gaps between scales. We demonstrate that LSTMs achieve reasonable performance in predicting hourly tree-level sap flow for different tree genera and in diverse climate zones and forest types. A key technical criteria for developing robust LSTMs include the random subsampling strategy to account for uncertainties in sap flow measurements and the splitting of the sap flow data into individual time series to enable model training across sites and measurement periods. Training LSTMs on large and diverse data sets and on several tree types proved beneficial compared to specialized models and supported the objective of using LSTMs to generalize vegetation dynamics beyond the individual tree-level sap flow measurement. By leveraging the capabilities of LSTMs to simulate state-dependent ecohydrological processes, we achieved generalization with minimal and broadly available forest characteristics and dynamic meteorological inputs. As more data sets become publicly accessible, we anticipate improvements in the precision and scope of such models, particularly for forest types underrepresented in SAPFLUXNET. Continental-scale LSTMs could significantly enhance our understanding of vegetation water use dynamics reflected in sap flow, potentially serving as critical benchmarks for partitioning evapotranspiration in land surface and hydrological models.

## Data Availability Statement

The data on which this article is based are available in Poyatos et al. (2021). All python codes to run the models are available in Loritz and Wu (2023).

## Acknowledgments

This research contributes to the ViTamins (Invigorating Hydrological Science and Teaching: merging key Legacies with new Concepts and Paradigms) Project funded by the Volkswagen Foundation. MB was supported by a U.S. Department of Agriculture NIFA Postdoctoral Fellowship (Grant 2023-67012-40086) and the LEMONTREE (Land Ecosystem Models based On New Theory, observations and Experiments) project, funded through the generosity of Eric and Wendy Schmidt by recommendation of the Schmidt Futures programme. Open Access funding enabled and organized by Projekt DEAL.

## References

- Bassiouni, M., Manzoni, S., & Vico, G. (2023). Optimal plant water use strategies explain soil moisture variability. *Advances in Water Resources*, 173, 104405. <https://doi.org/10.1016/j.advwatres.2023.104405>
- Besnard, S., Carvalhais, N., Altaf Arain, M., Black, A., Brede, B., Buchmann, N., et al. (2019). Memory effects of climate and vegetation affecting net ecosystem CO<sub>2</sub> fluxes in global forests. *PLoS One*, 14(2), e0211510. <https://doi.org/10.1371/journal.pone.0211510>
- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. Retrieved from <http://arxiv.org/abs/1406.1078>
- de Bartolomeis, P., Meterez, A., Shu, Z., & Stocker, B. D. (2023). An effective machine learning approach for predicting ecosystem CO<sub>2</sub> assimilation across space and time. <https://doi.org/10.5194/egusphere-2023-1826>
- Dugas, W. A., Heuer, M. L., Hunsaker, D., Kimball, B. A., Lewin, K. F., Nagy, J., & Johnson, M. (1994). Sap flow measurements of transpiration from cotton grown under ambient and enriched CO<sub>2</sub> concentrations. *Agricultural and Forest Meteorology*, 70(1–4), 231–245. [https://doi.org/10.1016/0168-1923\(94\)90060-4](https://doi.org/10.1016/0168-1923(94)90060-4)
- Dugas, W. A., Wallace, J. S., Allen, S. J., & Roberts, J. M. (1993). Heat balance, porometer, and deuterium estimates of transpiration from potted trees. *Agricultural and Forest Meteorology*, 64(1–2), 47–62. [https://doi.org/10.1016/0168-1923\(93\)90093-W](https://doi.org/10.1016/0168-1923(93)90093-W)
- Good, S. P., Noone, D., & Bowen, G. (2015). Hydrologic connectivity constrains partitioning of global terrestrial water fluxes. *Science*, 349(6244), 175–177. <https://doi.org/10.1126/science.aaa5931>
- Granier, A., & Loustau, D. (1994). Measuring and modelling the transpiration of a maritime pine canopy from sap-flow data. *Agricultural and Forest Meteorology*, 71(1–2), 61–81. [https://doi.org/10.1016/0168-1923\(94\)90100-7](https://doi.org/10.1016/0168-1923(94)90100-7)
- Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1–2), 80–91. <https://doi.org/10.1016/j.jhydrol.2009.08.003>
- Hassler, S. K., Weiler, M., & Blume, T. (2018). Tree-stand- and site-specific controls on landscape-scale patterns of transpiration. *Hydrology and Earth System Sciences*, 22(1), 13–30. <https://doi.org/10.5194/hess-22-13-2018>
- Hochreiter, S. (1998). Recurrent neural net learning and vanishing gradient. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(2), 107–116. <https://doi.org/10.1142/s0218488598000094>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Hoek van Dijke, A. J., Mallick, K., Teuling, A. J., Schlerf, M., Machwitz, M., Hassler, S. K., et al. (2019). Does the normalized difference vegetation index explain spatial and temporal variability in sap velocity in temperate forest ecosystems? *Hydrology and Earth System Sciences*, 23(4), 2077–2091. <https://doi.org/10.5194/hess-23-2077-2019>
- Jackisch, C., Knoblauch, S., Blume, T., Zehe, E., & Hassler, S. K. (2020). Estimates of tree root water uptake from soil moisture profile dynamics. *Biogeosciences*, 17(22), 5787–5808. <https://doi.org/10.5194/bg-17-5787-2020>
- Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., et al. (2022). Uncertainty estimation with deep learning for rainfall-runoff modeling. *Hydrology and Earth System Sciences*, 26(6), 1673–1693. <https://doi.org/10.5194/hess-26-1673-2022>
- Koppa, A., Rains, D., Hulsman, P., Poyatos, R., & Miralles, D. G. (2022). A deep learning-based hybrid model of global terrestrial evaporation. *Nature Communications*, 13(1), 1912. <https://doi.org/10.1038/s41467-022-29543-7>
- Köstner, B., Granier, A., & Cermák, J. (1998). Sapflow measurements in forest stands: Methods and uncertainties. *Annales Des Sciences Forestières*, 55(1–2), 13–27. <https://doi.org/10.1051/forest:19980102>
- Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, 23(12), 5089–5110. <https://doi.org/10.5194/hess-23-5089-2019>
- Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. In *Nature* (Vol. 521(7553), pp. 436–444). Nature Publishing Group. <https://doi.org/10.1038/nature14539>
- Li, Y., Ye, J., Xu, D., Zhou, G., & Feng, H. (2022). Prediction of sap flow with historical environmental factors based on deep learning technology. *Computers and Electronics in Agriculture*, 202. <https://doi.org/10.1016/j.compag.2022.107400>
- Loritz, R., Bassiouni, M., Hildebrandt, A., Hassler, S. K., & Zehe, E. (2022). Leveraging sap flow data in a catchment-scale hybrid model to improve soil moisture and transpiration estimates. *Hydrology and Earth System Sciences*, 26(18), 4757–4771. <https://doi.org/10.5194/hess-26-4757-2022>
- Loritz, R., Hassler, S. K., Jackisch, C., Allroggen, N., van Schaik, L., Wienhöfer, J., & Zehe, E. (2017). Picturing and modeling catchments by representative hillslopes. *Hydrology and Earth System Sciences*, 21(2), 1225–1249. <https://doi.org/10.5194/hess-21-1225-2017>
- Loritz, R., & Wu, C. (2023). Generalising tree-level sap flow across the European continent [Software]. *Zenodo*. <https://doi.org/10.5281/zenodo.10118262>
- Lundblad, M., Lagergren, F., & Lindroth, A. (2001). Evaluation of heat balance and heat dissipation methods for sapflow measurements in pine and spruce. *Annals of Forest Science*, 58(6), 625–638. <https://doi.org/10.1051/forest:2001150>
- Mai, J., Shen, H., Tolson, B. A., Gaborit, É., Arsenault, R., Craig, J. R., et al. (2022). The great lakes runoff intercomparison project phase 4: The great lakes (GRIP-GL). *Hydrology and Earth System Sciences*, 26(13), 3537–3572. <https://doi.org/10.5194/hess-26-3537-2022>
- Mai, H. R., Zheng, F., Gupta, H., Chen, J., Mai, J., Savic, D., et al. (2023). On how data are partitioned in model development and evaluation: Confronting the elephant in the room to enhance model generalization. *Environmental Modelling & Software*, 167, 105779. <https://doi.org/10.1016/j.envsoft.2023.105779>
- Poyatos, R., Granda, V., Flo, V., Adams, M. A., Adorján, B., Aguadé, D., et al. (2021). Global transpiration data from sap flow measurements: The SAPFLUXNET database [Dataset]. In *Earth System science data* (Vol. 13(6)). <https://doi.org/10.5194/essd-13-2607-2021>
- Renner, M., Hassler, S. K., Blume, T., Weiler, M., Hildebrandt, A., Guderle, M., et al. (2016). Dominant controls of transpiration along a hillslope transect inferred from ecohydrological measurements and thermodynamic limits. *Hydrology and Earth System Sciences*, 20(5), 2063–2083. <https://doi.org/10.5194/hess-20-2063-2016>
- Seeger, S., & Weiler, M. (2021). Temporal dynamics of tree xylem water isotopes: In situ monitoring and modeling. *Biogeosciences*, 18(15), 4603–4627. <https://doi.org/10.5194/bg-18-4603-2021>

- Singh, S. K., & Bárdossy, A. (2012). Calibration of hydrological models on hydrologically unusual events. *Advances in Water Resources*, 38, 81–91. <https://doi.org/10.1016/j.advwatres.2011.12.006>
- Steppe, K., de Pauw, D. J. W., Doody, T. M., & Teskey, R. O. (2010). A comparison of sap flux density using thermal dissipation, heat pulse velocity and heat field deformation methods. *Agricultural and Forest Meteorology*, 150(7–8), 1046–1056. <https://doi.org/10.1016/j.agrformet.2010.04.004>
- Tournebize, R., & Boistard, S. (1998). Comparison of two sap flow methods for the estimation of tree transpiration. *Annales Des Sciences Forestières*, 55(6), 707–713. <https://doi.org/10.1051/forest:19980605>
- Vandegehuchte, M. W., & Steppe, K. (2013). Corrigendum to: Sap-flux density measurement methods: Working principles and applicability. *Functional Plant Biology*, 40(10), 1088. [https://doi.org/10.1071/FP12233\\_CO](https://doi.org/10.1071/FP12233_CO)
- Wi, S., & Steinschneider, S. (2022). Assessing the physical realism of deep learning hydrologic model projections under climate change. *Water Resources Research*, 58(9), e2022WR032123. <https://doi.org/10.1029/2022WR032123>
- Zehe, E., Graeff, T., Morgner, M., Bauer, A., & Bronstert, A. (2010). Plot and field scale soil moisture dynamics and subsurface wetness control on runoff generation in a headwater in the Ore Mountains. *Hydrology and Earth System Sciences*, 14(6), 873–889. <https://doi.org/10.5194/hess-14-873-2010>