water

MDPI

Article

Introducing State-of-the-Art Deep Learning Technique for Gap-Filling of Eddy Covariance Crop Evapotranspiration Data

Lior Fine [1,2], Antoine Richard [3], Josef Tanny [1], Cedric Pradalier [3], Rafael Rosa [1] and Offer Rozenstein [1,*]

[1] Institute of Soil, Water and Environmental Sciences, Agricultural Research Organization—Volcani Institute, HaMaccabim Road 68, Rishon LeZion 75359, Israel; liorfine@gmail.com (L.F.); tanai@volcani.agri.gov.il (J.T.); rafael.rosa.uy@gmail.com (R.R.)

[2] Department of Soil and Water Sciences, Faculty of Agriculture, Food and Environment, The Hebrew University of Jerusalem, P.O. Box 12, Rehovot 76100, Israel

[3] GeorgiaTech Lorraine–UMI2958 GT-CNRS, 2 rue Marconi, 57057 Metz, France; antwoine31@gmail.com (A.R.); cedric.pradalier@georgiatech-metz.fr (C.P.)

* Correspondence: offerr@volcani.agri.gov.il

Abstract: Gaps often occur in eddy covariance flux measurements, leading to data loss and necessitating accurate gap-filling. Furthermore, gaps in evapotranspiration (ET) measurements of annual field crops are particularly challenging to fill because crops undergo rapid change over a short season. In this study, an innovative deep learning (DL) gap-filling method was tested on a database comprising six datasets from different crops (cotton, tomato, and wheat). For various gap scenarios, the performance of the method was compared with the common gap-filling technique, marginal distribution sampling (MDS), which is based on lookup tables. Furthermore, a predictor importance analysis was performed to evaluate the importance of the different meteorological inputs in estimating ET. On the half-hourly time scale, the deep learning method showed a significant 13.5% decrease in nRMSE (normalized root mean square error) throughout all datasets and gap durations. A substantially smaller standard deviation of mean nRMSE, compared with marginal distribution sampling, was also observed. On the whole-gap time scale (half a day to six days), average nMBE (normalized mean bias error) was similar to that of MDS, whereas standard deviation was improved. Using only air temperature and relative humidity as input variables provided an RMSE that was significantly smaller than that resulting from the MDS method. These results suggest that the deep learning method developed here is reliable and more consistent than the standard gap-filling method and thereby demonstrates the potential of advanced deep learning techniques for improving dynamic time series modeling.

Keywords: net radiation; air temperature; wind speed; relative humidity; time series

check for updates

Citation: Fine, L.; Richard, A.; Tanny, J.; Pradalier, C.; Rosa, R.; Rozenstein, O. Introducing State-of-the-Art Deep Learning Technique for Gap-Filling of Eddy Covariance Crop Evapotranspiration Data. *Water* **2022**, *14*, 763. https://doi.org/10.3390/w14050763

Academic Editor: Leonardo V. Noto

Received: 3 January 2022
Accepted: 24 February 2022
Published: 28 February 2022

1. Introduction

A precise estimation of crop evapotranspiration (ET) is of importance for quantifying terrestrial water budgets for irrigation purposes and for understanding evaporation, transpiration, and photosynthesis processes. Most ET estimation methods are indirect and thus provide only an approximate estimation. Direct methods are expensive and technically complex [1] but provide a fairly accurate and reliable estimate of ET. The eddy covariance (EC) method is a direct approach for measuring field-scale ET over crops [2]. Most EC systems provide a time series of half-hourly average fluxes of latent and sensible heat and CO_2 as well as momentum fluxes. Unfortunately, the percentage of missing data is between 20% and 60% of the original dataset [3] due to gaps of various lengths caused by different factors, including sensor malfunction, power breaks, and data quality filtering. To calculate daily, monthly, or yearly sums, these gaps must be reliably filled. Furthermore, when gaps occur in the data (e.g., when studying the ET diurnal curve or when comparing with other high-temporal-resolution ET measurement methods), accurate gap filling is required.

In order to solve the aforementioned problem, various gap-filling methods that use available data to reconstruct the missing parts have been developed [3–11]. Most of the methods are based on empirical techniques that derive and parameterize the relations between certain drivers and the fluxes. Usually, the drivers are meteorological variables (e.g., air temperature, solar radiation, etc.) measured on-site or at a nearby meteorological station.

Several methods have been suggested in the literature, from basic methods such as Mean Diurnal Course and Lookup Tables [4], and their integrated version (Mean Distribution Sampling; [5]), to more sophisticated ones such as nonlinear regression [3,4,6], artificial neural networks [7–9], and random forests [10]. A comprehensive comparison of 15 gap-filling techniques based on ten benchmark datasets showed that different methods performed almost equally well, suggesting little room for improvement [3]. However, these datasets are year-long and represent European forests, which are characterized by conditions that are much less time-varying than those of the rapidly changing croplands. Furthermore, most of the methods presented in that comparison and the broader literature have been developed and tested on CO_2 flux rather than on ET.

The EC method was also widely used to measure fluxes over various crops [12]. These crops are managed ecosystems with rapid change throughout a growing season. In a warm climate, as illustrated in this article, a typical crop cycle could last 3–4 months, during which the conditions in the field change dramatically, mainly due to variations in the leaf area index and canopy structure. As a result, the boundary layer properties and the albedo change during the cropping period, and large temporal variations of measured heat fluxes are observed [13]. Additionally, due to the warm climate, ET fluxes reach much higher values (and, as a result, higher errors when estimating them) than natural ecosystems such as the forests and grasslands in European climates. Furthermore, gaps occurring during the short growing season lead to a limited amount of valid data, putting a strain on modeling the data in the gaps. These characteristics of EC measurements over field crops make gap-filling a challenging task.

Only a few papers have dealt with gap-filling of crop ET—e.g., [4,14–16]. These studies used various methods—e.g., MDS, Kalman Filter, Multiple Linear Regression, and Mean Diurnal Variation (MDV). In recent years, however, new advances in machine learning, specifically deep learning (DL), were developed, with high capabilities in time series modeling. Nevertheless, so far, none of the studies on ET gap-filling available in the literature employed this approach. Hence, the major goal of this research was to examine the suitability of such a DL approach for filling gaps in eddy covariance ET data over crops.

The field of machine learning, also known as artificial intelligence, has been the focus of intense research during the past few decades, leading to significant breakthroughs in many fields. A subset of machine learning, deep learning, is based on artificial neural networks (ANNs). Unlike the 'classical' ANNs used in most studies on gap filling and modeling of EC flux data [7–9], deep learning models can learn to extract features and take advantage of the spatial or temporal structure of the data streams in a hierarchical way. EC and additional bio-meteorological measurements from flux towers generate multidimensional time series, and these, in general, have been the topic of application of these new classes of neural networks [17]. Recently, Reichstein et al. [18] highlighted the potential of using deep learning techniques in geoscience for modeling dynamic time series. In the context of ET estimation, deep learning has been evaluated for interpolating local ET prediction to regional scale by correlating terrain appearance with ET [19]. However, as far as we know, this kind of model has never been examined in the context of gap-filling of EC data. Recently, we proposed a model for gap filling based on attention networks that account for the periodicity within the data [11]. In the current paper, we put this model to the test using real-world EC data to evaluate its performance under several scenarios detailed below.

This research aimed to evaluate deep neural networks for gap-filling of EC measurements over crops, assuming they could explain the temporal structure of the data, thereby overcoming the obstacles arising from a limited amount of data with rapid temporal changes. Furthermore, specific objectives of this study included: (1) comparing the model performances on different gap durations; (2) testing the effect of adding a non-irrigated crop (wheat) to a training set based on irrigated crops and gap-filled data from irrigated crops (tomato and cotton); and (3) assessing the outcome of using four meteorological variables (net radiation, air temperature, relative humidity, and wind speed) and a subset of them as input values on the performance of the neural networks. The findings could help scientists using the EC system to measure fluxes of ET and may highlight the potential of using these deep learning techniques for interpolating other cyclic multivariate time series.

2. Materials and Methods

2.1. ET Datasets

This study is based on six datasets of EC measurements over tomato, cotton, and wheat crops in two regions in Israel. Tomato and cotton crops were measured in the Hula Valley (northern Israel; 33°08′ N, 35°36′ E), while wheat crops were located in the central-south coastal plain (31°39′ N, 34°37′ E) (Figure 1, Table 1). Both regions have a Mediterranean climate: The summer is hot, rainless, and predominantly sunny with little variation from day to day, and winter is cool and rainy. Spring wheat, in Israel, is sown in the autumn and is generally a non-irrigated crop. On the other hand, tomato and cotton are grown in the dry summer and rely entirely on irrigation. These differences affect the range of ET values of the different crops as a result of water availability and meteorological conditions (see ET values in Figure 2, meteorological characteristics in Table 2, and typical diurnal courses in Figure 3).

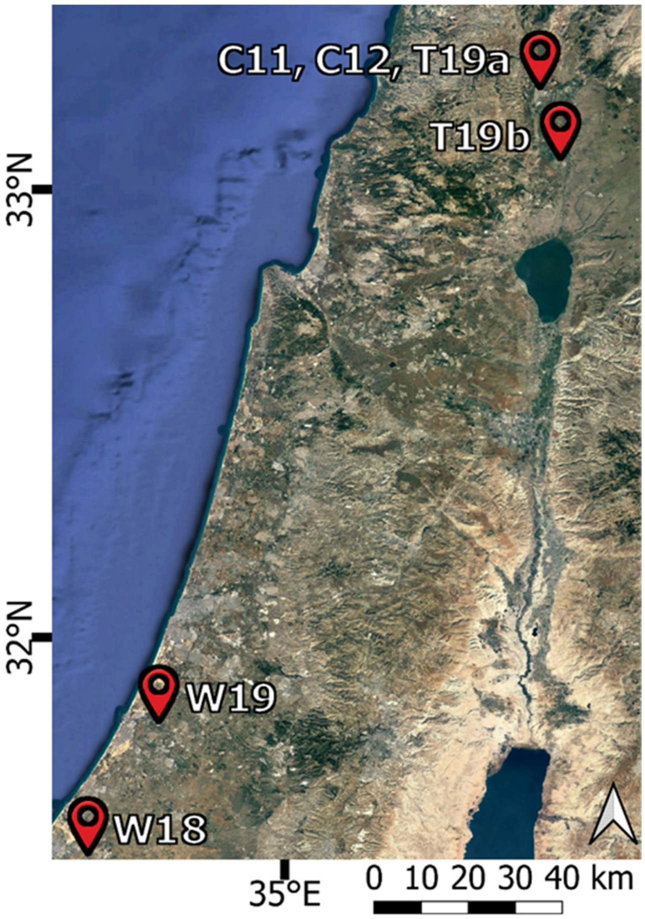

Figure 1. Map of northern-central Israel showing the locations of the different evapotranspiration measurement sites, as indicated in Table 1.

Table 1. Details on datasets used to train and evaluate the suggested deep learning gap-filling method.

#	Site	Region	Crop	Dates	No. of Full Days
1	C11	Hula Valley	Cotton	1 June 2011–5 September 2011	64
2	C12	Hula Valley	Cotton	1 June 2012–12 September 2012	96
3	W18	Coastal Plain	Wheat	10 January 2018–8 April 2018	65
4	W19	Coastal Plain	Wheat	21 December 2018–11 April 2019	94
5	T19a	Hula Valley	Tomato	2 May 2019–25 July 2019	83
6	T19b	Hula Valley	Tomato	24 April 2019–15 August 2019	92

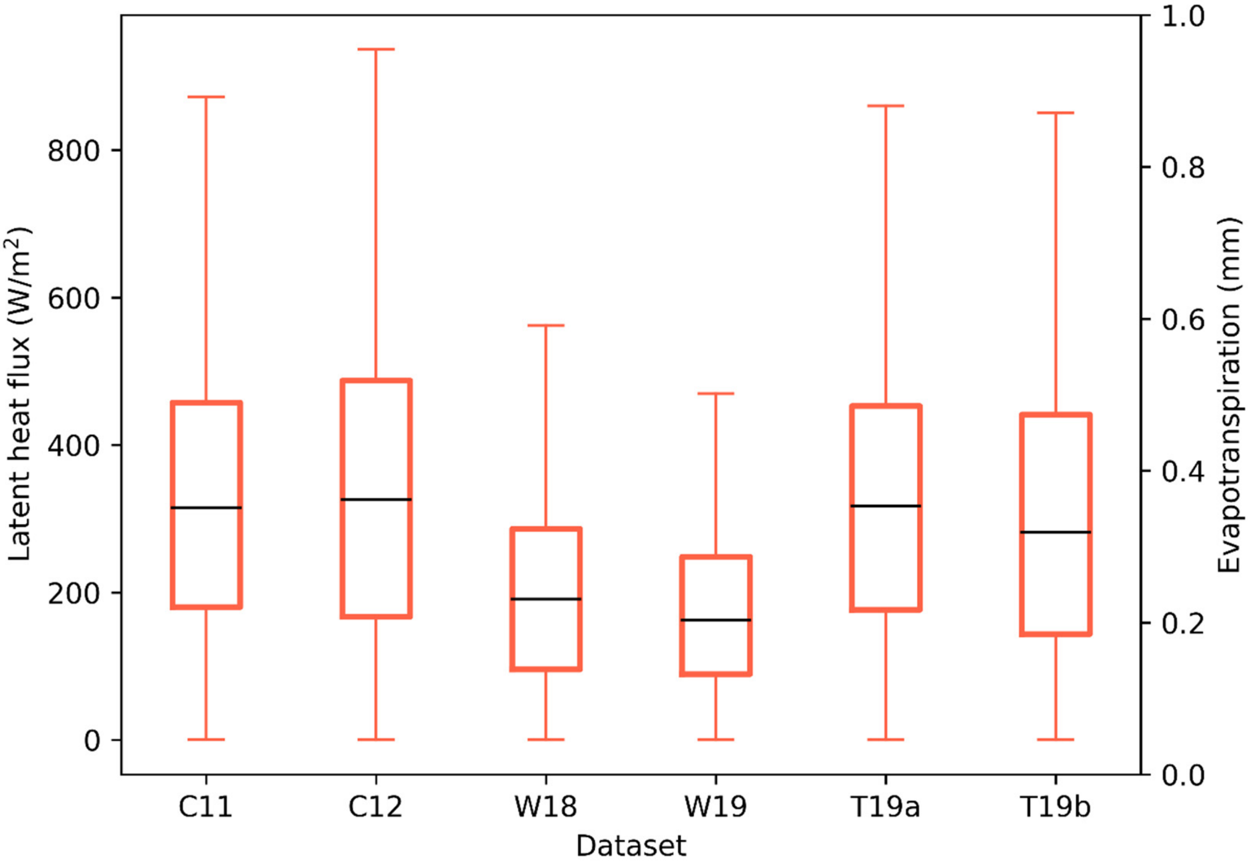

Figure 2. Latent heat flux and evapotranspiration half-hourly values during daytime (net radiation > 0) in the different datasets. The box extends from the lower to upper quartile values of the data, with a median line. The whiskers extend from the box to show the range of the data. 'C' stands for cotton, 'W' for wheat, and 'T' for tomato. For further details, see Table 1.

Table 2. Meteorological characteristics of the datasets used to train and evaluate the suggested deep learning gap-filling method. The values are the median of the half-hourly means; the 5th and 95th percentiles are contained in the brackets.

#	Site	Air Temperature (°C)	Relative Humidity (%)	Daily Average Net Radiation (W/m^2)	Wind Speed (m/s)
1	C11	25.6 (18.4, 32.7)	71.6 (45.8, 95.7)	178.4 (151.3, 198.0)	0.6 (0.1, 4.8)
2	C12	27.1 (19.3, 34.4)	64.2 (35.0, 97.6)	173.2 (149.0, 187.8)	0.9 (0.2, 5.8)
3	W18	15.8 (10.2, 23.7)	75.0 (38.0, 95.1)	111.6 (44.8, 176.9)	1.2 (0.4, 4.8)
4	W19	13.8 (7.4, 20.0)	81.4 (49.8, 95.9)	82.5 (33.0, 152.5)	1.7 (0.3, 4.7)
5	T19a	26.3 (16.5, 35.4)	60.7 (24.2, 86.7)	173.1 (147.2, 206.5)	1.1 (0.2, 6.1)
6	T19b	26.6 (16.7, 35.9)	63.1 (26.2, 90.7)	186.4 (122.2, 198.3)	1.1 (0.2, 4.4)

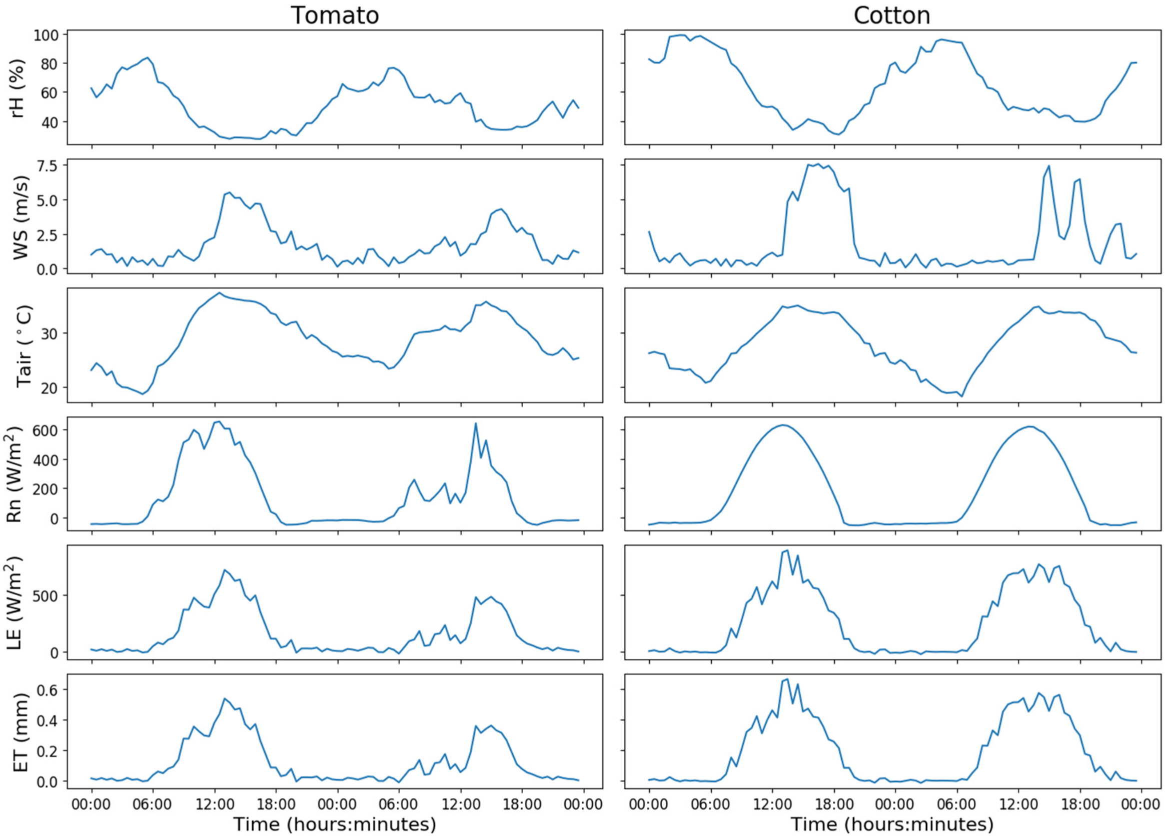

Figure 3. Diurnal courses of latent heat flux, evapotranspiration, and the four meteorological inputs for two typical days in the tomato crops (**left**) and cotton crops (**right**). rH: relative humidity, WS: wind speed, Tair: air temperature, Rn: net radiation, LE: latent heat flux, ET: evapotranspiration.

Each dataset consisted of half-hourly latent heat flux (LE; associated with ET) along with meteorological variables measured at the same tower: net radiation (Rn; W·m^{-2}), air temperature (Ta; °C), relative humidity (rH; %), and wind speed (WS; m·s^{-1}). These variables were chosen because, except for Rn, they are prevalent in meteorological stations. Accordingly, even if there were no data from the local system (e.g., due to a power outage), it would still be possible to fill the gaps. The datasets consisted of raw 20 Hz data acquired by EC systems consisting of an open-path gas analyzer (LI-7500DS; LI-COR Biosciences, Lincoln, NE, USA) and a three-axis ultrasonic anemometer (Model CSAT3; Campbell Scientific, Logan, UT, USA). In addition to the EC system, the experimental setup of each campaign included a net radiometer (Q*7.1; REBS, Seattle, WA, USA) and soil heat flux plates (HFT-3.1; REBS, Seattle, WA, USA). Air temperature and humidity were measured by a combined temperature–humidity sensor (HMP45, Campbell Sci., Logan, UT, USA). All study sites were flat, horizontal, and uniformly managed crop fields. In each field, the Eddy Covariance sensors were positioned both above the roughness sub-layer and within the constant-flux surface layer. This was verified by applying a standard height/fetch ratio analysis on the cotton fields, or by a detailed footprint analysis using the Kljun et al. [20] model on the remaining fields. These analyses confirmed that in all campaigns, the 90% flux footprint was within the field under study.

Processing of the cotton datasets, C11 and C12, included coordinate system rotation [21], path averaging, frequency–response corrections [22], and density correction [23]. Further details on datasets C11 and C12 can be found in Rosa and Tanny [13]. The four other datasets raw data of the EC systems and meteorological variables were processed using EddyPro©, version 7.0 (LI-COR Biosciences, USA). The procedure employed by EddyPro included linear detrending, correction of low-pass [24] and high-pass [25] filtering effects, covariance maximization, and density fluctuation compensation with the Webb correction [23]. Spikes, amplitude resolution artifacts, and unrealistic drop-outs, including other artifacts, were filtered using the statistical tests of Vickers and Mahrt [26]. To measure the evapotranspiration, our study relied on an open-path infrared gas analyzer whose measurements are affected by rain. Therefore, to avoid measurement error, the rainy days were filtered out of the dataset. Across all six datasets, we only filtered out seven rainy days, mostly from the wheat datasets. Analysis of data quality based on the fulfillment of the theoretical assumptions of EC method showed that the data that needed to be filtered were mostly at night (only ~2% of poor data were found during daytime), when latent heat flux values were close to zero, and thus negligible. Since the deep neural network developed here requires that the data for the training process will be continuous in time, and due to their negligible contribution to the flux, these low-quality data points were not filtered out. However, the final filtering of the EC, as well as of the corresponding meteorological data, was completed by meticulous inspection to detect and remove points of clear error. Finally, we filled short gaps using linear interpolation, while days containing longer gaps during daytime were filtered out entirely, and the neighboring days were joined together.

2.2. Deep Learning Architecture

Artificial neural networks (ANNs) build a nonlinear function to describe a mapping between a set of inputs (or 'predictors') and an output ('target'). In this study, a deep neural network based on a multi-head attention [11,27] was used. This network architecture leverages information from both past, present, and future. The type of architecture we chose is commonly referred to as a sequence-to-sequence model. Formally, our problem can be formulated as in Equation (1), where $x_i(t)$ is the set of observed variables such that $\forall i \in [1,n]$, $\forall t \in [1,T]$, $x_i(t) \in \mathbb{R}$ and $z_i(t)$ is the set of non-observed variables such that $\forall i \in [1,m]$, $\forall t \in [1,T]$, $x_i(t) \in \mathbb{R}$.

$$y(t) = f(x_1(t), \ldots, x_n(t), z_1(t), \ldots, z_n(t)) \in \mathbb{R}$$

$$\forall i \in [1,n],\ \exists \tau_i \in \mathbb{R}\ s.t.\ x_i(t) = x_i(t+\tau_i) + \varepsilon \tag{1}$$

The 'sequence of observations' will be noted as x and the 'sequence of targets' as y. The model used here aims at recovering the missing values from the target sequence using both the sequence of observations and the target sequence with missing values. The artificial neural network used here is inspired from Transformer networks [25]. Unlike recurrent ANNs, Transformers do not iteratively process all the elements of a sequence. Instead, they process the whole of the sequence at once. This allows them to be more efficient and more performant than their recurrent counterparts. In [11], which we use in this work, they remove the decoder of the original architecture to retain only the encoder. This simplification of the encoder/decoder architecture effectively reduces the number of parameters within the model, reducing its complexity and making it easier to train. Similar to Transformer, their architecture features multiple attention-heads that process embedded input.

An attention-head (or Scaled Dot-Product Attention) computes the correlation for every combination of element pairs in the sequence under the form of an attention matrix. This propagates information between two elements, even if they are far apart in the sequence. Multi-head attention consists in using multiple attention-heads in parallel (with the same input) and then concatenates their output. In a regular Transformer, the network processes the whole of the variables at once, input sequence and target sequence together. However, in our case, the model should not use information from the variables inside the gap of the target sequence. At the same time, it must leverage the information from the input sequence to fill the gap. Hence, they manually assigned different variables to each network head. They used three attention heads: one head processes the observation sequence, another head processes the target sequence, and finally, the last head processes a concatenation of observation and target sequences. This reduces the learning complexity of the model and helps it perform better in low data-regimes.

Another key difference with regular Transformers is that they use a cyclical positional encoding. In Transformers, the positional encoding gives a unique value (or identifier) to each element of the sequence allowing the model to know the position of each element in the sequence. The cyclicity in the positional encoding allows the model to account for the periodic time dependencies natively. They rely on the same equation as [27], and do not learn this encoding. This makes the learning process simpler, which is desirable regarding the limited amount of training examples at our disposal.

2.3. Dataset Processing

To train the model, we first divided the database into training, validation, and test sets. The training set was used to fit the parameters of the model. Subsequently, the validation set was used to evaluate the predictions of the model with the fitted parameters, and finally, the test set provided an unbiased evaluation of the final model. At first, splitting each dataset into the different sets was considered, but this turned out to be impractical: First, the training, validation, and test sets must be representative for the entire duration of the season and that would have created very fragmented data which would have made it challenging to train the model. Second, this meant that the training process, which requires knowledge, computing power, and time, had to be performed by the end-user (instead of the preferred situation where the end-user receives a trained model from a maintained database). To solve the above problems, we decided to split the whole database (all six seasons from Table 1) into a training dataset consisting of the data of five seasons and a second dataset (of the sixth season) that would be divided into validation and test sets (see Figure 4). The sixth dataset was divided in a way that would be representative of the entire season: For short and medium gaps, the dataset was divided into ten pieces (five validation sets and five test sets), and for long gaps, it was split into six pieces (see definition of gap lengths in Section 2.4.1). Finally, we applied a cross-validation technique, switching between all possible combinations of datasets (seasons) included in training and validation-test sets. We used these datasets as they were acquired on different crops during different seasons. Hence, by testing our model on a given dataset and training on the rest, we ensured that they would generalize well to similar crops under unforeseen conditions.

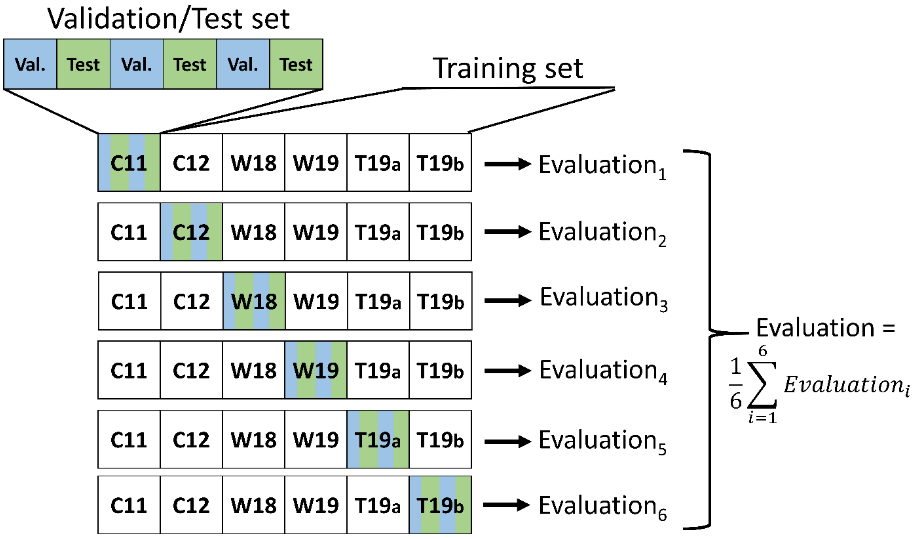

Figure 4. Cross-validation scheme. In each evaluation, a different site was used to derive the validation/test set (site abbreviations according to Table 1).

2.4. Analyses and Performance Evaluation

2.4.1. Comparison with MDS

The marginal distribution sampling (MDS) method was used to compare the performance with the deep learning (DL) model developed here because it is currently one of the most widely used gap-filling methods. MDS is based on lookup tables and mean diurnal course [4,5]. In short, within a certain time window, the missing value is replaced by the average value under similar meteorological conditions from preceding and succeeding days (i.e., with a lookup table). If the meteorological conditions are not similar enough within the starting time window of seven days, the window size is increased [28]. The present DL model is compared against a gap-filling tool based on MDS that was developed at the Max Planck Institute for Biogeochemistry in Jena and is available as an R package called REddyProc [29]. Filling the lookup table cells requires defining a range of allowed deviation for each variable, and we used net radiation, relative humidity, and air temperature ranges of 50 W m^{-2}, 10%, and 1.5 °C, respectively, as suggested by Foltýnová et al. [16].

We compared our method with MDS on multiple scenarios: short gaps, medium gaps, and long gaps. The exact gap length for each was randomly chosen: short (24 to 64 half-hourly data points or half-day to 32 h), medium (64 to 144 points or 32 h to three days), and long (144 to 288 points or three to six days). We assumed gaps shorter than half a day would easily be filled, even when using more straightforward methods such as linear interpolation, and therefore, they were not examined. Longer gaps were not considered in this study due to limitations in our model setting. Furthermore, as mentioned in Section 2.3, a cross-validation technique was used, so that the DL method performance on each dataset was evaluated by a trained model based on the other five datasets. DL and MDS methods were both tested only on the test datasets. The number of repetitions was 5400 (three trained models * six datasets * three gaps * 100 gaps).

2.4.2. Training the Model with or without Wheat Datasets

In this test, our database was divided into two groups: irrigated summer crops (cotton and tomato) and non-irrigated spring crops (wheat). The two groups have different characteristics that are reflected in different averages and ranges of latent heat flux values (Figure 2), mostly derived from different weather conditions (Table 2) and irrigation management. We tested how adding or removing the non-irrigated datasets (i.e., wheat datasets) from the training set affected the model performances on the irrigated crops (i.e., tomato and cotton datasets). To test this, we trained a model on three irrigated crop datasets and cross-validated it on the fourth irrigated field. The two non-irrigated crops were then added to the training set, and the new model was cross-validated again on the irrigated crops. The performances were evaluated using the measures in Equations (2)–(6) (Section 2.5) but only for medium gap lengths (day and a half to three days). The number of repetitions was 1200 (three trained models * four datasets * 100 gaps).

2.4.3. Sensitivity Analysis of Predictors

To examine the effect of different predictors on model performance, we trained models based on all possible combinations of one, two, three, or all four predictors (net radiation, air temperature, relative humidity, and wind speed). In addition, a model that does not use meteorological variables at all and depends entirely on the time of day was trained. Subsequently, MDS was applied using the full set of variables, and performances were evaluated according to the measures in Equations (2)–(6) (Section 2.5) but only for medium gap lengths (day and a half to three days). The number of repetitions was 1800 (three trained models * six datasets * 100 gaps), for each combination of variables.

2.5. Evaluation Protocol and Metrics

To compare the performances of the different approaches we used the following procedure: For each repetition of performance evaluation on a dataset, we generated a randomly located artificial gap with a total of 100 gaps for each repetition. The performance of a method for each dataset was evaluated by comparing each individual estimated (filled) LE value with the observed LE value.

The statistical performance measures used in the analysis included root mean square error (RMSE; Equation (2)) to evaluate the error on each half-hourly point, and mean bias error (MBE; Equation (3)) to illustrate the bias induced on the total duration of the gap, ranging from half a day to six days. MBE is a measure relevant for analyzing periodic flux sums (e.g., daily sums). A normalized version of the two metrics (nRMSE and nMBE; Equations (4) and (5)) was used to account for the different LE averages of the datasets (Figure 2). In addition, we calculated the Nash–Sutcliffe model efficiency coefficient (NSE; Equation (6)) as in Nash and Sutcliffe [30].

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum(p_i - o_i)^2} \tag{2}$$

$$\mathrm{MBE} = \frac{1}{n}\sum(p_i - o_i) \tag{3}$$

$$\mathrm{nRMSE} = \frac{\mathrm{RMSE}}{\overline{y}} \tag{4}$$

$$\mathrm{nMBE} = \frac{\mathrm{MBE}}{\overline{y}} \tag{5}$$

$$\mathrm{NSE} = 1 - \frac{\sum(p_i - o_i)^2}{\sum(o_i - \overline{o})^2} \tag{6}$$

where p_i denotes the individual estimated LE value, o_i the observed (measured) value, $\overline{o}$ the mean of observed values, n the number of observations, and $\overline{y}$ the mean LE value of each dataset.

In applied machine learning, each training of the same DL algorithm is called a 'model'. Because the learning process is stochastic, each training will result in unique parameters, and therefore the predictions will be different. Thus, when evaluating ANNs, the results are always performed on three models trained with different random seeds. This ensures that the results presented here are reproducible and not acquired from a single cherry-picked model. Finally, the statistical significance was tested using ANOVA and pairwise comparisons using Tukey's HSD test (α = 0.01).

3. Results

3.1. Comparison with MDS

Figure 5 shows an example of diurnal courses of LE, as measured by the EC system and estimated by the DL and MDS approaches from the 'T19a' campaign (see Table 1). A 6-day artificial gap was applied, as shown in Figure 5a, while Figure 5b illustrates the data with the filled gap. As the figure shows, the suggested DL method was successfully able to fill gaps up to six days long. Analysis throughout all datasets and gap lengths showed that the DL approach, on average, had significantly lower RMSE and higher NSE when compared with MDS (Figure 6). MBE was not significantly different, but its standard deviation with the DL method was lower (Figure 6).

After comparing the performance of our models across all datasets and gap lengths, we studied their performances on each dataset. In this scenario, each model was tested on a specific dataset, while being trained on the other five datasets. The MDS was operated only on the tested dataset. The DL was more accurate than the MDS throughout all datasets, as demonstrated by a significantly higher decrease in mean nRMSE, ranging between 3.5% and 36.4% (Figure 7). The standard deviation of DL was also lower than that of the MDS in all datasets, except for 'C12′ (Figure 7). The superior performance of the DL models over the MDS for datasets it has never encountered before can be explained by its ability to extract local trends. These models are trained to fill the measurement not only by using the meteorological variable but also by using the values surrounding them. Using these values, the model can scale its predictions and hence can achieve high-quality results even on new datasets that it has never encountered before.

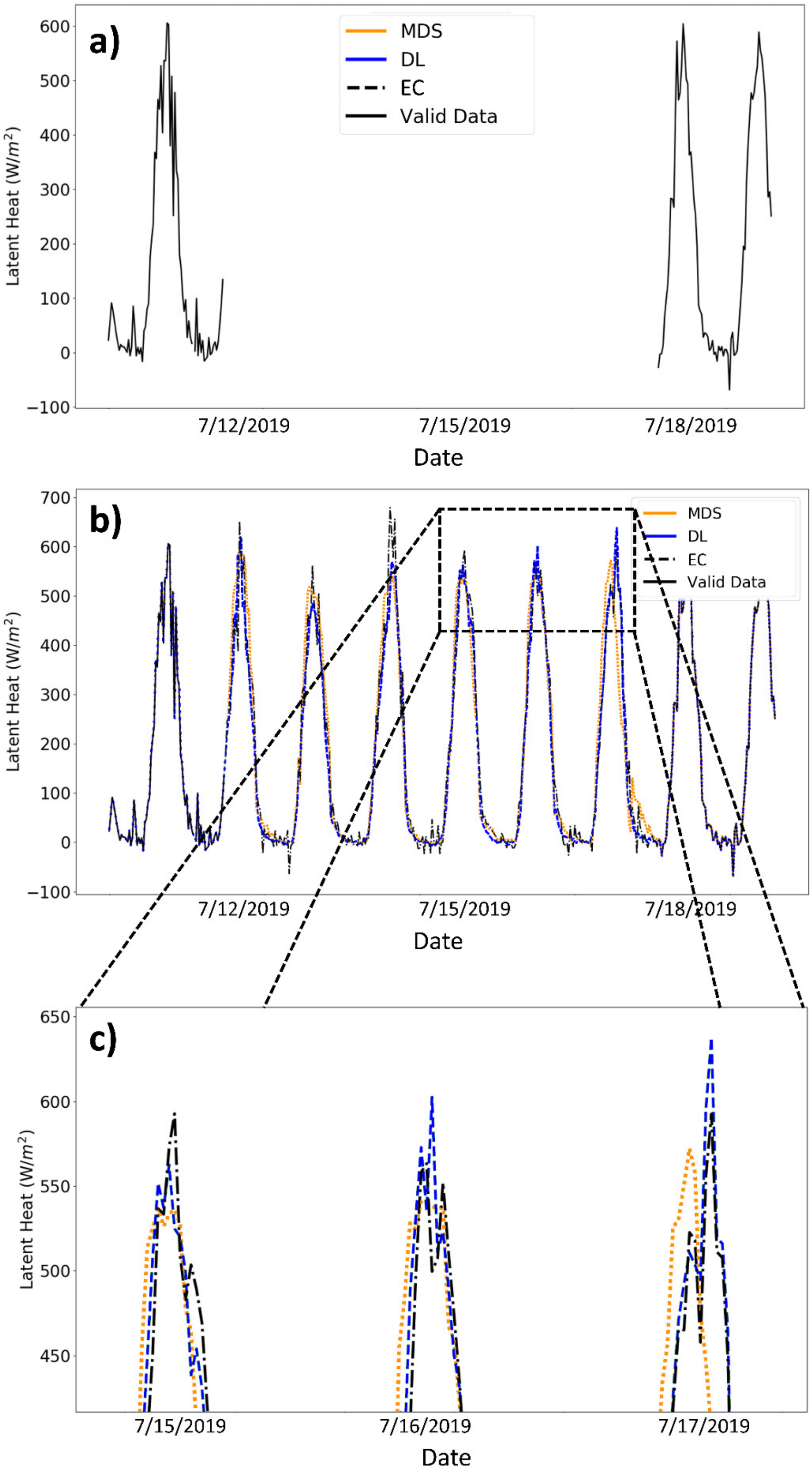

Figure 5. An example of gap-filling in latent heat flux data of tomato in the Hula Valley, Israel (11 July 2019–19 July 2019), with an artificial gap of six days (288 points). 'Valid Data' is the sequence that the deep learning (DL) method is exposed to at the moment of gap-filling. In the above, (**a**) denotes a sequence with an artificial 6-day gap; (**b**) the results of filling the gap using 'deep learning' (DL) and 'marginal distribution sampling' (MDS) compared with the measured eddy covariance sequence (EC); and (**c**) zoom into the results.

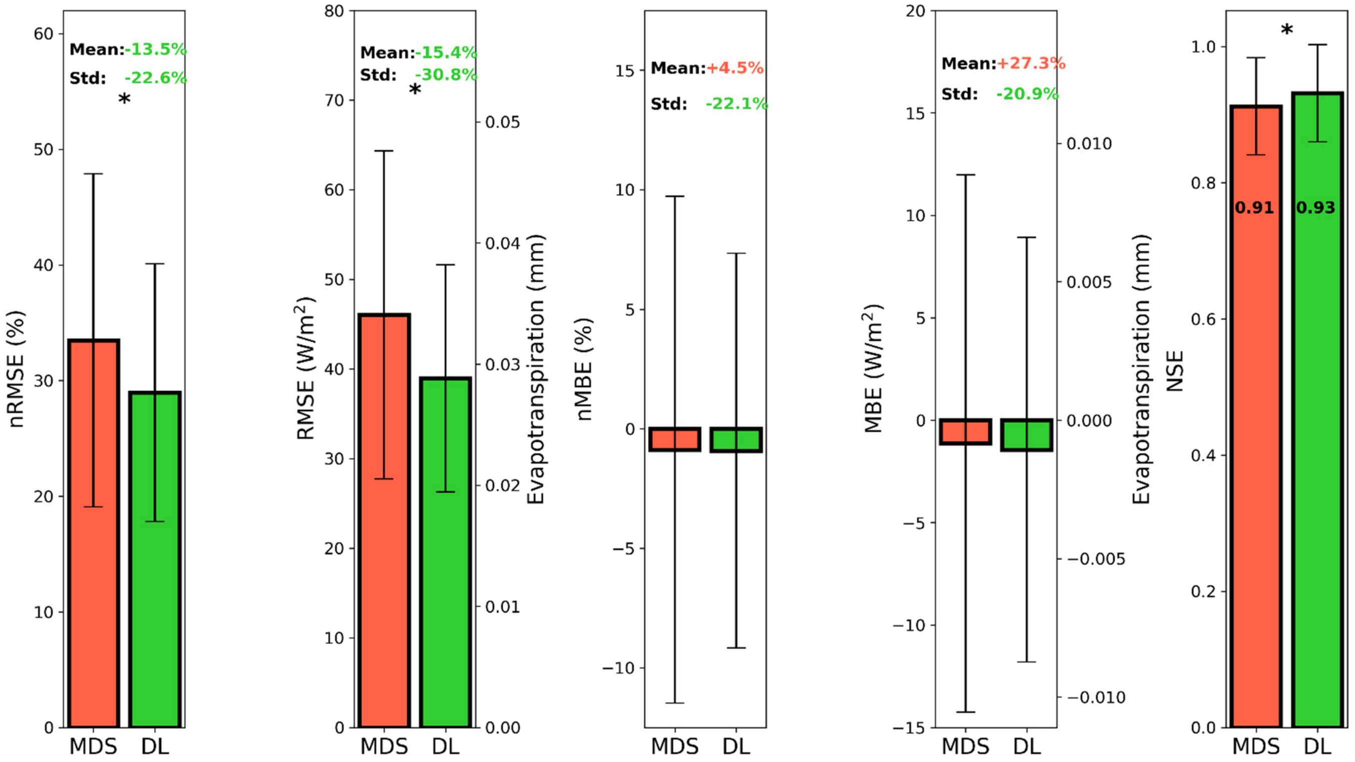

Figure 6. A comparison between our suggested deep learning gap-filling method (DL) and the common method, marginal distribution sampling (MDS), throughout all datasets and gap lengths. Green font color indicates an improvement in performance (of DL over MDS), orange indicates that MDS performed better than DL. Bars represent the standard deviation of the mean (n = 5400). Asterisks indicate statistically significant differences ($p < 0.01$). Negative values indicate a reduction, i.e., Std:−5% means that the DL has a 5% smaller standard deviation than MDS. RMSE and MBE are presented as normalized values (nRMSE and nMBE, in percentage (%)) as well as absolute values in Wm^{-2} and mm.

We then compared the performance of the DL and MDS on different gap sizes. Figure 8 summarizes this experiment. Across all gap sizes, the DL performance was better than the MDS performance. This is demonstrated by the significantly lower mean RMSE obtained by the DL (Figure 8). The decrease in RMSE of DL in comparison with MDS was significantly larger for short gaps than for medium and long gaps; no significant difference between DL performance for medium and long gaps was found.

Adding the two wheat datasets for training in addition to the four summer crops set did not produce a significant difference in the performances of DL on the irrigated crops (tomato and cotton), with minimal change in the RMSE: 24.65 W m^{-2} with the wheat and 24.72 W m^{-2} without.

3.2. Sensitivity Analysis of Predictors

Figure 9 shows the performance of the DL model with various combinations of the input variables. Excluding net radiation from the model inputs (‘rH-Tair-WS’ in Figure 9) caused a significant decrease in performance, while leaving in Rn but excluding any other variables (air temperature, wind speed, or relative humidity) did not result in a significant reduction in performance compared with the full set of variables (Figure 9). This correlation of variables relative to the quality of prediction was expected and demonstrates that our models follow the same rules as classical modelling techniques. Using only air temperature and relative humidity as predictors provided lower RMSE than the MDS model. Using a single predictor, net radiation was the best predictor among the meteorological inputs, and DL performance was significantly better than MDS performance with all four input variables. Using no meteorological variables (Figure 9: W/o met. vars.)—i.e., depending only on the time of day—showed the most unsatisfactory performance but was not significantly different from MDS performance. This can be explained by the ability of the DL model to predict correct-looking sequences following neighboring patterns. However, if the weather conditions are radically different in the gap, the model will fail to reconstruct it accurately.

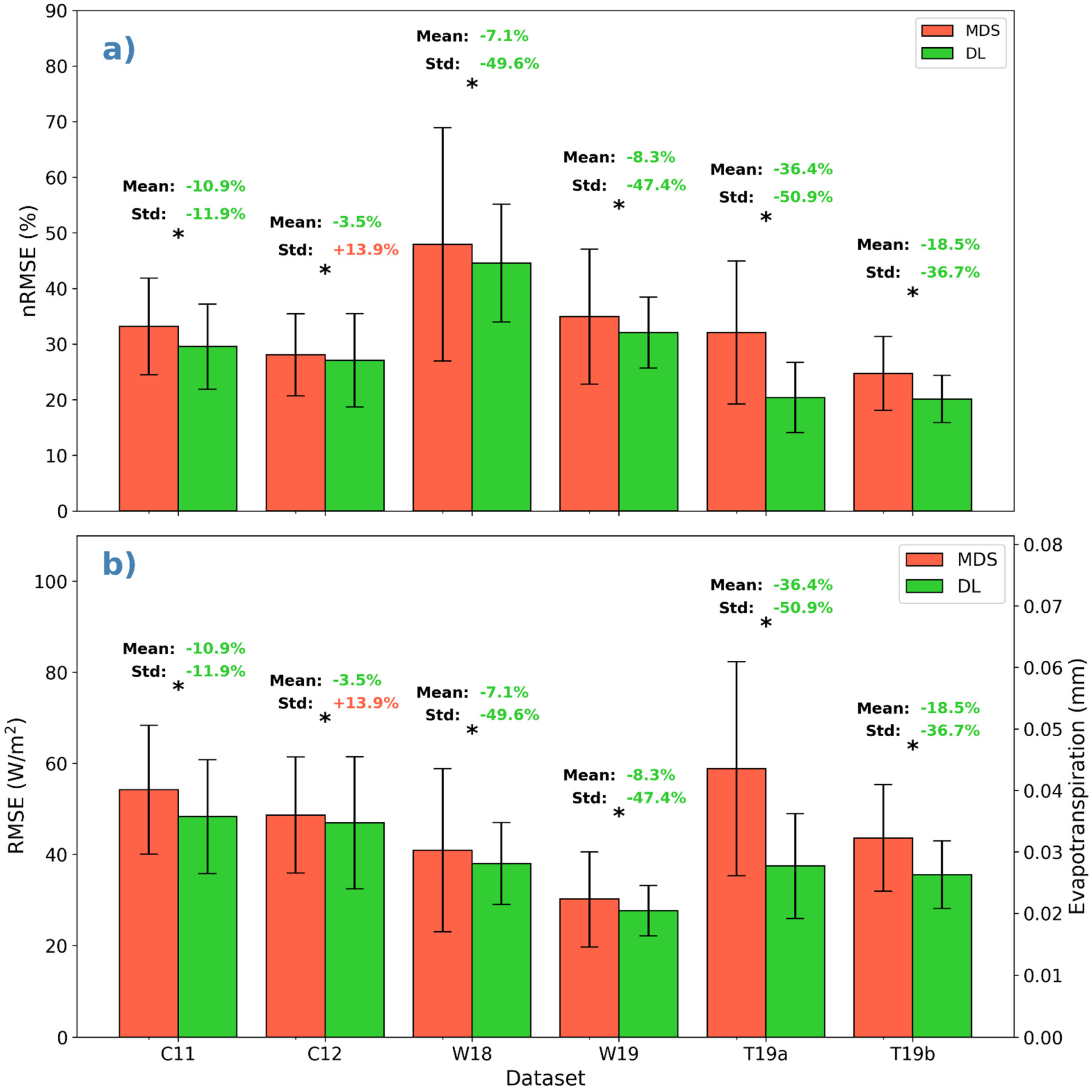

Figure 7. Comparing the present deep learning gap-filling method (DL) and the common method, marginal distribution sampling (MDS), testing different datasets: (**a**) normalized RMSE; (**b**) RMSE of half hourly latent heat flux or the half-hourly ET (mm). Green font color indicates an improvement in performance (of DL over MDS), and orange indicates a decline. Bars represent the standard deviation of the mean (n = 900). Asterisks indicate statistically significant differences ($p < 0.01$). Negative values indicate a reduction, i.e., Std: −5% means that the DL has a 5% smaller standard deviation than MDS.

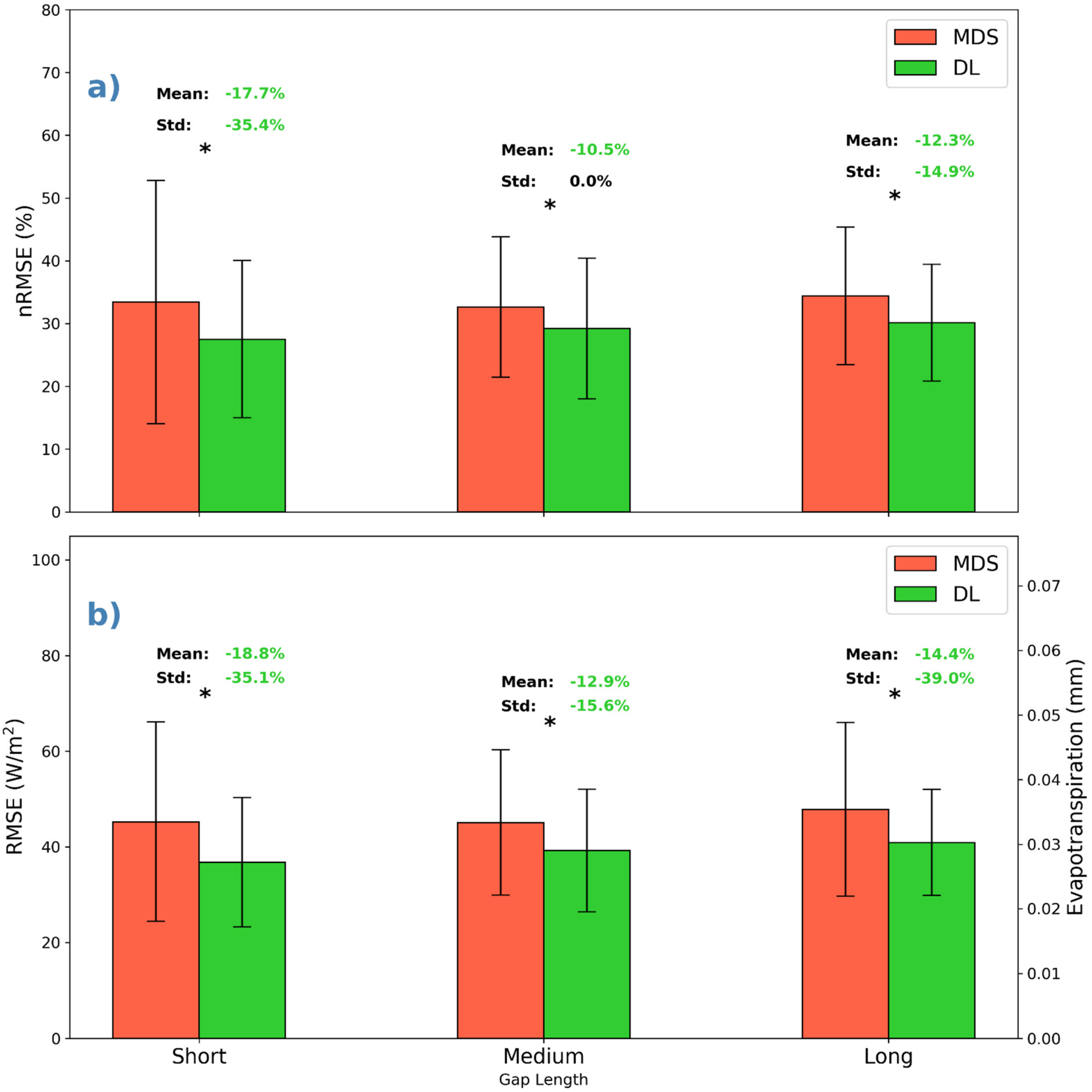

Figure 8. A comparison between the deep learning gap-filling method (DL) and the common method, marginal distribution sampling (MDS), testing different gap lengths: Short = half-day to 32 h; medium = 32 h to three days; and long = three to six days. (**a**) Normalized RMSE; (**b**) RMSE of half hourly latent heat flux or the half-hourly ET (mm). Bars represent the standard deviation of the mean (n = 1800). Asterisks indicate statistically significant differences ($p < 0.01$). Negative values indicate a reduction, i.e., Std: −5% means that the DL has a 5% smaller standard deviation than MDS.

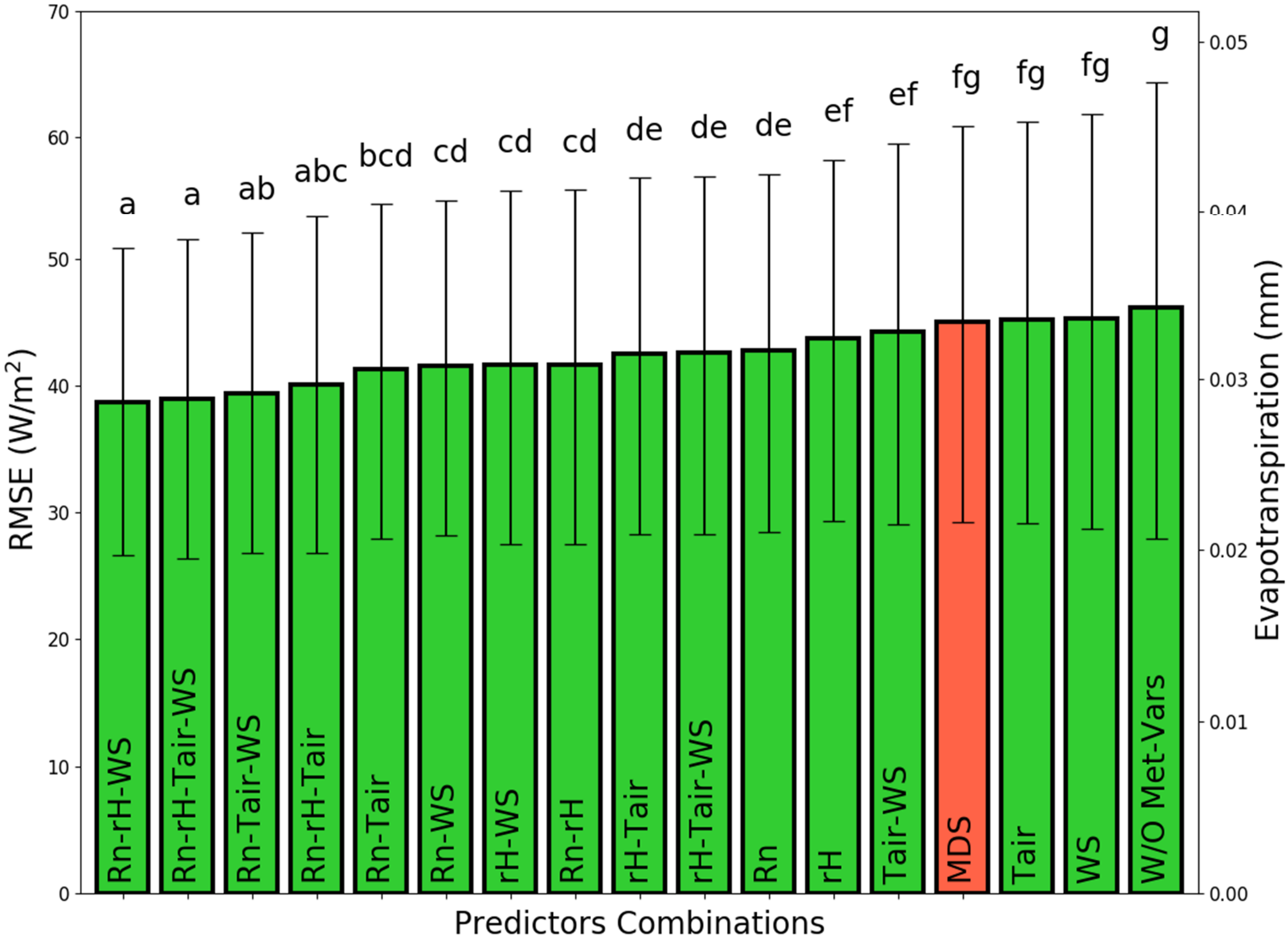

Figure 9. Deep learning (DL) model performances using combinations of different meteorological variables (Rn—net radiation; rH—relative humidity; WS—wind speed; Tair—air temperature; 'W/o met. vars.'—without meteorological variables, i.e., time of day only) as model inputs compared with the common gap-filling method, marginal distribution sampling (MDS). Bars represent the standard deviation of the mean (n = 1800). Different letters indicate statistically significant differences between groups (p < 0.01).

4. Discussion

In this study, we examined a new deep learning gap-filling method for crop evapotranspiration measured by eddy covariance. The method was tested on a database containing different crops and demonstrated superior performance than the widely used gap-filling method, MDS. Overall, the DL method showed a significant decrease in RMSE throughout all datasets and gap lengths (Figure 6). Furthermore, the DL method had a lower standard deviation (22% less) in both nRMSE and nMBE. Models trained using all four available meteorological variables as predictors, and models missing either air temperature, wind speed, or relative humidity, showed the best performance of the method. On the contrary, excluding net radiation impaired performance significantly. Moreover, including wheat datasets in the training set and evaluating the method on tomato and cotton datasets did not improve or reduce performance.

The DL method examined here was able to 'learn' from a relatively small database containing only six ET measurement campaigns over agricultural fields and was able to successfully fill gaps of up to six days. Across all different datasets and gap lengths, the DL method showed a more accurate (lower RMSE) and precise (lower standard deviation) prediction of ET flux than did MDS. These results are in line with [3], which showed around 12% improvement in RMSE using ANNs compared with MDS. However, the present DL method has some crucial advantages over these ANNs. First, using a model developed for a certain crop and season, the DL method does not need to be trained on the specific dataset being gap-filled. Secondly, the DL includes natural embedding of time—i.e., the method learns the data as a time sequence rather than as individual points. The benefits are (1) faster runtime, on the order of seconds versus minutes; (2) easier implementation after the model is trained (these two points are due to the way the DL was implemented, see Section 2.2); and (3) minimal data input of about 12 days instead of a full-year, as required for the conventional ANN [3], making the method more suitable for short growing seasons of annual crops. The minimal data input is possible because our DL model relies mainly on the values of the neighboring days and, in addition, on modeling of the relations between the meteorological variables and the flux. The ANN methods [3] require a whole year because they are based on modeling the relationship between the meteorological variables and the flux only, without an understanding of the time or the sequence of the values. Therefore, the DL approach is advantageous compared with ANN, especially for cases with limited training data.

Mean nMBE was small and similar in both DL and MDS ($\sim -1\%$), but the standard deviation of nMBE in the DL method was lower by 22.1% than that of MDS. The small nMBE means that the bias induced on the gap-filled sequences was very low for both methods, but the DL was more consistent. These results indicate that on the time scale of the whole gap (half a day to six days), the performance of the DL and MDS methods in calculating the total flux was the same. However, the suggested DL method can lower the uncertainty of the prediction.

Regarding the performance for each dataset individually, the errors introduced by the DL method were consistently significantly lower than those of MDS, and the decrease in nRMSE ranged between 3.5% and 36.4% (Figure 7). This means that the method performed well on six different datasets and three different crops, illustrating the reliability of the method. Testing for different gap lengths (Figure 8) showed that RMSE of DL was significantly lower than that of MDS and that it was consistent throughout all gap lengths.

Another important step was testing the effect in the training set of including a crop with different characteristics than that of the crop that needs to be gap-filled. Including and omitting the wheat datasets from training on a dataset that included the irrigated crops and gap-filling the tomato and cotton datasets did not result in a significant difference in performance. This may indicate that the method is insensitive to adding data from different field crops and therefore allows for extending the database that will be used to train the models.

Performing a sensitivity analysis on the predictors showed an increase in RMSE when net radiation was excluded. Excluding any other variable did not result in a significant increase in RMSE. Furthermore, using net radiation as a single predictor showed a significantly lower RMSE than any other single variable, even when comparing with the MDS method. The high correlation between latent heat flux and net radiation in relation to the other meteorological variables can be seen in the diurnal courses (Figure 3), where LE and Rn curves are usually aligned well. These results indicate that net radiation is the most important predictor of ET, relating to the fact that radiation is the primary source of energy for the ET process, and therefore it is the limiting factor in the system. This is in agreement with Ambas and Baltas [31], who showed that solar radiation was the most essential factor when tested on different reference ET calculation methods. Additionally, they showed that air temperature was the second most important factor, although this is not evident in this study, perhaps due to auto-correlation of the data (the similarity between points at the same time of day but on neighboring days), which reduces the importance of the predictors, in general. Additionally and interestingly, using the DL model that was trained using only the time of the day as input without any meteorological variable as a predictor showed no significant difference in RMSE compared with the MDS method. This may be due to the importance of time dependency within the DL algorithm, while this factor is not embedded in the MDS method.

Some limitations of the suggested gap-filling method should be noted. Although this study suggests that the method is insensitive to adding data from different crops, the model must be trained on data similar enough to that of the gap-filled dataset (similar crops, climate, ecosystems, etc.) because of its dependency on previously trained datasets. Additionally, the method was tested on a relatively small database with specific characteristics, and more testing under different conditions and in different ecosystems and ET fluxes is needed.

The model suggested here is based on recent advances in deep learning methodologies [11]. Hence, as is, it cannot be used at this stage as a convenient off-the-shelf tool for routine gap-filling like the MDS tool developed by Reichstein et al. [29].

5. Conclusions

A novel DL technique was demonstrated to be an appropriate solution for gap-filling of eddy covariance ET flux data of annual field crops in Israel. The high accuracy at the half-hourly time-scale makes the method especially valuable when accurate high-temporal resolution data are necessary. Additionally, on a daily time-scale, the method seems to lower the uncertainty of the ET estimation but does not improve the mean bias induced compared with MDS. Under the conditions examined here, the method appears to be insensitive to gap size. In addition, the performance on crops of a particular type (irrigated crops in this case) does not seem to be impaired by adding a crop with different characteristics to the training set. This is an example of the use of state-of-the-art deep learning networks to better understand dynamic micrometeorological processes.

It would be interesting to train this kind of method on a much larger database, including varying fluxes, ecosystems, and conditions. With a more extensive database, the DL method may show even better performance because of the nature of machine learning, which typically improves when large datasets are available for training.

Author Contributions: Conceptualization, O.R., C.P. and J.T.; methodology, L.F., R.R. and A.R.; software, L.F. and A.R.; validation, L.F., R.R. and A.R.; formal analysis, L.F.; writing—original draft preparation, L.F., J.T. and O.R.; writing—review and editing, L.F., J.T., A.R., C.P. and O.R.; visualization, L.F.; supervision, O.R., J.T. and C.P.; project administration, O.R. and C.P.; funding acquisition, O.R., C.P. and J.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by a grant from the Ministry of Science & Technology, Israel (grant number 3-15830); the Ministry of Europe and Foreign Affairs (MEAE); and the Ministry of Higher Education, Research and Innovation (MESRI) of France.

Data Availability Statement: Weather data is publicly available at: https://meteo.co.il/ (accessed on 23 February 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Weksler, S.; Rozenstein, O.; Haish, N.; Moshelion, M.; Walach, R.; Ben-Dor, E. A hyperspectral-physiological phenomics system: Measuring diurnal transpiration rates and diurnal reflectance. *Remote Sens.* **2020**, *12*, 1493. [CrossRef]
2. Aubinet, M.; Vesala, T.; Papale, D. *Eddy Covariance: A Practical Guide to Measurement and Data Analysis*; Springer Science & Business Media: Berlin, Germany, 2012.
3. Moffat, A.M.; Papale, D.; Reichstein, M.; Hollinger, D.Y.; Richardson, A.D.; Barr, A.G.; Beckstein, C.; Braswell, B.H.; Churkina, G.; Desai, A.R. Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. *Agric. For. Meteorol.* **2007**, *147*, 209–232. [CrossRef]
4. Falge, E.; Baldocchi, D.; Olson, R.; Anthoni, P.; Aubinet, M.; Bernhofer, C.; Burba, G.; Ceulemans, R.; Clement, R.; Dolman, H. Gap filling strategies for long term energy flux data sets. *Agric. For. Meteorol.* **2001**, *107*, 71–77. [CrossRef]
5. Reichstein, M.; Falge, E.; Baldocchi, D.; Papale, D.; Aubinet, M.; Berbigier, P.; Bernhofer, C.; Buchmann, N.; Gilmanov, T.; Granier, A. On the separation of net ecosystem exchange into assimilation and ecosystem respiration: Review and improved algorithm. *Glob. Chang. Biol.* **2005**, *11*, 1424–1439. [CrossRef]
6. Lloyd, J.; Taylor, J.A. On the temperature dependence of soil respiration. *Funct. Ecol.* **1994**, *8*, 315–323. [CrossRef]

7. Papale, D.; Valentini, R. A new assessment of European forests carbon exchanges by eddy fluxes and artificial neural network spatialization. *Glob. Chang. Biol.* **2003**, *9*, 525–535. [CrossRef]
8. Braswell, B.H.; Sacks, W.J.; Linder, E.; Schimel, D.S. Estimating diurnal to annual ecosystem parameters by synthesis of a carbon flux model with eddy covariance net ecosystem exchange observations. *Glob. Chang. Biol.* **2005**, *11*, 335–355. [CrossRef]
9. Moffat, A.M. A New Methodology to Interpret High Resolution Measurements of Net Carbon Fluxes Between Terrestrial Ecosystems and the Atmosphere. Ph.D. Thesis, Friedrich Schiller University, Jena, Germany, 2010.
10. Kim, Y.; Johnson, M.S.; Knox, S.H.; Black, T.A.; Dalmagro, H.J.; Kang, M.; Kim, J.; Baldocchi, D. Gap-filling approaches for eddy covariance methane fluxes: A comparison of three machine learning algorithms and a traditional method with principal component analysis. *Glob. Chang. Biol.* **2020**, *26*, 1499–1518. [CrossRef]
11. Richard, A.; Fine, L.; Rozenstein, O.; Tanny, J.; Geist, M.; Pradalier, C. Filling Gaps in Micro-Meteorological Data. In *European Conference on Machine Learning*; Springer: Cham, Switzerland, 2020.
12. Moureaux, C.; Ceschia, E.; Arriga, N.; Béziat, P.; Eugster, W.; Kutsch, W.L.; Pattey, E. Eddy covariance measurements over crops. In *Eddy Covariance*; Springer: Dordrecht, The Netherlands, 2012; pp. 319–331.
13. Rosa, R.; Tanny, J. Surface renewal and eddy covariance measurements of sensible and latent heat fluxes of cotton during two growing seasons. *Biosyst. Eng.* **2015**, *136*, 149–161. [CrossRef]
14. Alavi, N.; Warland, J.S.; Berg, A.A. Filling gaps in evapotranspiration measurements for water budget studies: Evaluation of a Kalman filtering approach. *Agric. For. Meteorol.* **2006**, *141*, 57–66. [CrossRef]
15. Boudhina, N.; Zitouna-Chebbi, R.; Mekki, I.; Jacob, F.; Ben Mechlia, N.; Masmoudi, M.; Prévot, L. Evaluating four gap-filling methods for eddy covariance measurements of evapotranspiration over hilly crop fields. *Geosci. Instrum. Methods Data Syst.* **2018**, *7*, 151–167. [CrossRef]
16. Foltýnová, L.; Fischer, M.; McGloin, R.P. Recommendations for gap-filling eddy covariance latent heat flux measurements using marginal distribution sampling. *Theor. Appl. Climatol.* **2020**, *139*, 677–688. [CrossRef]
17. Längkvist, M.; Karlsson, L.; Loutfi, A. A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognit. Lett.* **2014**, *42*, 11–24. [CrossRef]
18. Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N. Deep learning and process understanding for data-driven Earth system science. *Nature* **2019**, *566*, 195–204. [CrossRef] [PubMed]
19. Xu, T.; Guo, Z.; Liu, S.; He, X.; Meng, Y.; Xu, Z.; Xia, Y.; Xiao, J.; Zhang, Y.; Ma, Y. Evaluating different machine learning methods for upscaling evapotranspiration from flux towers to the regional scale. *J. Geophys. Res. Atmos.* **2018**, *123*, 8674–8690. [CrossRef]
20. Kljun, N.; Calanca, P.; Rotach, M.; Schmid, H. A simple two-dimensional parameterisation for Flux Footprint Prediction (FFP). *Geosci. Model Dev.* **2015**, *8*, 3695–3713. [CrossRef]
21. Kowalski, A.S.; Anthoni, P.M.; Vong, R.J.; Delany, A.C.; Maclean, G.D. Deployment and evaluation of a system for ground-based measurement of cloud liquid water turbulent fluxes. *J. Atmos. Ocean. Technol.* **1997**, *14*, 468–479. [CrossRef]
22. Moore, C.J. Frequency response corrections for eddy correlation systems. *Bound.-Layer Meteorol.* **1986**, *37*, 17–35. [CrossRef]
23. Webb, E.K.; Pearman, G.I.; Leuning, R. Correction of flux measurements for density effects due to heat and water vapour transfer. *Q. J. R. Meteorol. Soc.* **1980**, *106*, 85–100. [CrossRef]
24. Moncrieff, J.B.; Massheder, J.M.; De Bruin, H.; Elbers, J.; Friborg, T.; Heusinkveld, B.; Kabat, P.; Scott, S.; Soegaard, H.; Verhoef, A. A system to measure surface fluxes of momentum, sensible heat, water vapour and carbon dioxide. *J. Hydrol.* **1997**, *188*, 589–611. [CrossRef]
25. Moncrieff, J.; Clement, R.; Finnigan, J.; Meyers, T. Averaging, detrending, and filtering of eddy covariance time series. In *Handbook of Micrometeorology*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 7–31.
26. Vickers, D.; Mahrt, L. Quality control and flux sampling problems for tower and aircraft data. *J. Atmos. Ocean. Technol.* **1997**, *14*, 512–526. [CrossRef]
27. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
28. Wutzler, T.; Lucas-Moffat, A.; Migliavacca, M.; Knauer, J.; Sickel, K.; Šigut, L.; Menzer, O.; Reichstein, M. Basic and extensible post-processing of eddy covariance flux data with REddyProc. *Biogeosciences* **2018**, *15*, 5015–5030. [CrossRef]
29. Reichstein, M.; Moffat, A.M.; Wutzler, T.; Sickel, K. REddyProc: Data Processing and Plotting Utilities of (Half-) Hourly Eddy-Covariance Measurements. R Package Version 0.6–0/r9. 2014, p. 755. Available online: https://www.bgc-jena.mpg.de/bgi/index.php/Services/REddyProcWeb (accessed on 23 February 2022).
30. Nash, J.E.; Sutcliffe, J.V. River flow forecasting through conceptual models, I: A discussion of principles. *J. Hydrol.* **1970**, *10*, 398–409. [CrossRef]
31. Ambas, V.T.; Baltas, E. Sensitivity analysis of different evapotranspiration methods using a new sensitivity coefficient. *Glob. NEST J.* **2012**, *14*, 335–343.