A comparison of gap-filling algorithms for eddy covariance fluxes and their drivers

Atbin Mahabbati[1], Jason Beringer[1], Matthias Leopold[1], Ian McHugh[2], James Cleverly[3], Peter Isaac[4], Azizallah Izady[5]

[1] School of Agriculture and Environment, The University of Western Australia, 35 Stirling Hwy, Crawley, Perth WA, 6009, Australia

[2] School of Ecosystem and Forest Sciences, The University of Melbourne, Richmond, VIC, 3121, Australia

[3] School of Life Sciences University of Technology Sydney Broadway NSW 2007
[4] OzFlux Central Node, TERN Ecosystem Processes, Melbourne, VIC 3159, Australia
[5] Water Research Center, Sultan Qaboos University, Muscat, Oman

Correspondence to: Atbin Mahabbati (atbin.m@hotmail.com)

Abstract

The errors and uncertainties associated with gap-filling algorithms of water, carbon and energy fluxes data, have always been one of the main challenges of the global network of microclimatological tower sites that use eddy covariance (EC) technique. To address these concerns, and find more efficient gap-filling algorithms, we reviewed eight algorithms to estimate missing values of environmental drivers, and separately, nine algorithms for the three major fluxes in EC time series. We then examined the algorithms' performance for different gap-filling scenarios utilising the data from five EC towers during 2013. This research's objectives were a) to evaluate the impact of the gap lengths on the performance of each algorithm; b) to compare the performance of traditional and new gap-filling techniques for the EC data, for fluxes and separately for their corresponding meteorological drivers. The algorithms' performance was evaluated by generating nine gap windows with different lengths, ranging from a day to 365 days. In each scenario, a gap period was chosen randomly, and the data were removed from the dataset, accordingly. After running each scenario, a variety of statistical metrics were used to evaluate the algorithms' performance. The algorithms showed different levels of sensitivity to the gap lengths; The Prophet Forecast Model (FBP) revealed the most sensitivity, whilst the performance of artificial neural networks (ANNs), for instance, did not vary as much by changing the gap length. The algorithms' performance generally decreased with increasing the gap length, yet the differences were not significant for the windows smaller than 30 days. No significant difference between the algorithms was recognised for the meteorological and environmental drivers. However, the linear algorithms showed slight superiority over those of machine learning (ML), except the random forest algorithm estimating the ground heat flux (RMSEs of 28.91 and 33.92 for RF and CLR respectively). However, for the major fluxes, ML

Formatted: Justify Low

algorithms and the MDS showed superiority over the other algorithms. Even though ANNs, random forest (RF) and extreme gradient boost (XGB) showed comparable performance in gap-filling of the major fluxes, RF provided more consistent results with slightly less bias, as against the other ML algorithms. The results indicated that there is no single algorithm which outperforms in all situations, but the RF is a potential alternative for the ANNs as regards flux gap-filling.

1. Introduction

To address the global challenges of climatological and ecological changes, environmental scientists and policymakers are demanding data that are continuous in time and space. Besides, there is a need for quantifying and reducing uncertainties in such data, including observations of carbon, water and energy exchanges that are crucial components in national/international flux networks and global earth observing systems. Satellites partially fill this gap as they provide excellent spatial coverage but at a limited temporal resolution, and not measured at a point scale. As such, high-quality long-term site observations of ecosystem process and fluxes are needed that are continuous in time and space. The global eddy covariance (EC) flux tower networks (FLUXNET), consisted of its regional counterparts (i.e. AmeriFlux, EUROFLUX, OzFlux, etc.), was established in the late 1990s to address the global demand for such information (Aubinet et al., 1999; Baldocchi et al., 2001; Beringer et al., 2016a; Hollinger et al., 1999; Menzer et al., 2013; Tenhunen et al., 1998). Despite EC data being frequently used to validate process modelling analyses, field surveys and remote sensing assessments (Hagen et al., 2006), there are some serious concerns regarding the challenges associated with the technique, e.g. data gaps and uncertainties. Hence, filling data gaps and reducing uncertainties through better gap-filling techniques are highly needed.

Even though the EC is a common technique to measure fluxes of carbon, water and energy, there are some challenges in providing robust, high-quality continuous observations. One of the challenges regarding the technique, and therefore, the flux networks, is addressing data gaps and the uncertainties associated with the gap-filling process, mainly when the gap windows are long (longer than 12 consecutive days, as described by (Moffat et al., 2007)). These gaps happen very often due to a variety of reasons, such as values out of range, spike detection or manual exclusion of date and time ranges, instrument or power failure, herbivores, fire, eagles nests, cows, lightning, researchers on leave, etc. (Beringer et al., 2016b). Since EC flux towers are often located in harsh climates, their data are more susceptible to adverse weather (i.e. rain conditions), and they sometimes prevent quick access to sites for repair and maintenance. As a result, this issue can, in turn, produce gaps which might be relatively long (Isaac et al., 2017), and thus, problematic as follows. Firstly, loss of data is considered a threat to scientific studies depending on the missing data quantity, pattern, mechanism and nature (Altman and Bland, 2007; Molenberghs et al., 2014; Tannenbaum, 2010). That is because using an incomplete dataset might lead to biased, invalid and unreliable results (Allison, 2000; Kang, 2013; Little, 2002). Second, continuous gap-filled data are required to calculate the annual or monthly budgets of carbon or water balance components (Hutley et al., 2005).

Deleted: The errors and uncertainties associated with gap-filling algorithms of water, carbon and energy fluxes data, have always been one of the prominent challenges of the global network of microclimatological tower sites that use eddy covariance (EC) technique. To address this concern, and find more efficient gap-filling algorithms, we reviewed eight algorithms to estimate missing values of environmental drivers, and separately three major fluxes in EC time series. We then examined the performance of mentioned algorithms for different gap-filling scenarios utilising data from five OzFlux Network towers during 2013. The objectives of this research were a) to evaluate the impact of training and testing window lengths on the performance of each algorithm; b) to compare the performance of traditional and new gap-filling techniques for the EC data, for fluxes and their corresponding meteorological drivers. The performance of algorithms was evaluated by generating nine different training-testing window lengths, ranging from a day to 365 days. In each scenario, the gaps covered the data for the entirety of 2013 by consecutively repeating them, where, in each step, values were modelled by using earlier window data. After running each scenario, a variety of statistical metrics was used to evaluate the performance of the algorithms. The algorithms showed different levels of sensitivity to training-testing windows; The Prophet Forecast Model (FBP) revealed the most sensitivity, whilst the performance of artificial neural networks (ANNs), for instance, did not vary considerably by changing the window length. The performance of the algorithms generally decreased with increasing training-testing window length, yet the differences were not considerable for the windows smaller than 60 days. Gap-filling of the environmental drivers showed there was not a significant difference amongst the algorithms, the linear algorithms showed slight superiority over those of machine learning (ML), except the random forest algorithm estimating the ground heat flux (RMSEs of 30.17 and 34.93 for RF and CLR respectively). For the major fluxes, though, ML algorithms showed superiority (9 % less RMSE on average), except the Support Vector Regression (SVR), which provided significant bias in its estimations. Eve ...

Deleted: ¶

Deleted: the

Formatted: Font: (Default) Palatino Linotype, 10.5 pt, Complex Script Font: +Headings CS (Times New Roman), 10.5 pt

Deleted: Despite the capability of EC to frequently validate process modelling analyses

Other than the challenges caused by missing data, there are several sources of errors and uncertainties in the EC technique. Firstly, random error is associated with the stochastic nature of turbulence, associated sampling errors (incomplete sampling of large eddies, uncertainty in the calculated covariance between the vertical wind velocity and the scalar of interest), instrument errors, and footprint variability (Aubinet et al., 2012a). For instance, Dragoni et al. (2007) analysed an EC-based data of Morgan-Monroe State Forest for eight years (1999-2006) and assessed that instrument uncertainty was equal to 3 % of the total annual NEE. Another primary source of uncertainty in EC measurements is systematic errors that are usually caused by methodological challenges and instrument calibration problems (e.g. sonic anemometer errors, spikes, gas analyser errors, etc.). Finally, one of the sources of uncertainties is data processing, especially data gap-filling (Isaac et al., 2017; Moffat et al., 2007; Richardson et al., 2012; Richardson and Hollinger, 2007).

There are several uncertainties pertaining to gap-filling of missing values, including measurement uncertainty (Richardson and Hollinger, 2007), lengths and timing the gaps (Falge et al., 2001; Richardson and Hollinger, 2007) and the particular gap-filling algorithm that is used (Falge et al., 2001; Moffat et al., 2007). However, there are two dominant issues of long data gaps and the choice of a particular gap-filling algorithm (Aubinet et al., 2012a). Firstly, long gaps can significantly increase the total amount of uncertainty as the ecosystem behaviour might change because of different agricultural periods or phenological phases (e.g. growing season, harvest period, bushfire, etc.). And thereby show different responses under similar meteorological conditions (Aubinet et al., 2012a; Isaac et al., 2017; Richardson and Hollinger, 2007). Consequently, the period in which a long gap happens is essential. For example, research undertook by Richardson & Hollinger (2007) on data from a range of FLUXNET sites revealed that a week data gap during spring green-up in a forest led to a higher uncertainty over a three-week gap period during winter. Second, each gap-filling algorithm has its strengths and weaknesses; for instance, Moffat et al. (2007) compared 15 different commonly-used gap-filling algorithms. They found that there was not a significant difference between the performances of the algorithms with "good" reliability based on analysis of variance of RMSE. Besides, the overall gap-filling uncertainty was within ±25 g C m^{-2} yr^{-1} for most of the proper algorithms, whereas, the other algorithms generated higher uncertainties of up to ±75 g C m^{-2} yr^{-1}, showing that the uncertainty provided by reliable methods can be considerably smaller. This result is similar to the findings of Richardson & Hollinger (2007) who found as for the datasets used in the study, uncertainties of up to ±30 g C m^{-2} yr^{-1} for long gaps by appropriate algorithms. Considering that the data provided by EC tower networks are of use for research, government and policymakers, robust gap-filling is a need to quantify and reduce uncertainties in flux estimations.

Deleted: a couple of

Deleted: .

To manage the missing data problem, several methods have been typically used to fill data gaps in both fluxes and their meteorological drivers. Due to computational constraints of complex algorithms, early works to impute EC data gaps used interpolation methods based mostly on linear regression or temporal autocorrelation (Falge et al., 2001; Lee et al., 1999). These approaches were replaced quickly by more sophisticated methods such as non-linear regressions (Barr et al., 2004; Falge et al., 2001; Moffat et al., 2007; Richardson et al., 2006); lookup tables (Falge et al., 2001; Law et al., 2002; Zhao and Huang, 2015); artificial neural networks (ANNs) (Aubinet et al., 1999; Beringer et al., 2016a; Cleverly et al., 2013; Hagen et al., 2006; Isaac et al., 2017; Kunwor et al., 2017; Moffat et al., 2007; Papale and Valentini, 2003; Pilegaard et al., 2001; Staebler, 1999); mean diurnal variation (Falge et al., 2001; Moffat et al., 2007; Zhao and Huang, 2015), multiple imputations (Hui et al., 2004; Moffat et al., 2007), etc. Each of these methods has its pros and cons as follows: a) Interpolation methods such as the Mean Diurnal Variation (MDV), do not need any drivers, yet, their accuracy is lower than other approaches (Aubinet et al., 2012a). Moreover, this method may provide biased results on extremely clear or cloudy days (Falge et al., 2001). MDV is not recommended when a gap is longer than two weeks, for it cannot consider the non-linear relations between the drivers and the flux, and thus leads to a high level of uncertainty (Falge et al., 2001). And b) The Lookup table, especially its modified version, Marginal Distribution Sampling (MDS), has provided performance close to ANNs, and are more reliable and consistent than the other algorithms so far. Hence, MDS was chosen as one of the standard gap-filling methods in EUROFLUX (Aubinet et al., 2012a). Nevertheless, one of the concerns regarding this algorithm is that the independent variables, here meteorological drivers, might be auto-correlated. c) ANNs have commonly been used to gap-fill EC fluxes since 2000 and because of their robust and consistent results are considered as a standard gap-filling algorithm in several networks, e.g. ICOS, FLUXNET, OzFlux, etc. (Aubinet et al., 2012a; Beringer et al., 2017; Isaac et al., 2017). Despite their reliable performance, ANNs –and generally all other ML algorithms- face some challenges. Over-fitting, for instance, is a big concern and can happen when the number of degrees of freedom is high, while the training window is not long enough respectively, or the quality of the training dataset is low. This challenge becomes acute when the gaps happen within a period when the ecosystem behaviour is changing and thereby showing different response under similar meteorological conditions. Furthermore, there is a desire to have the training windows short so that the algorithm can track the ecosystem behaviour shift. Yet, this increases the risk of over-fitting depending on the algorithm. In other words, the training window length should be neither too short to cause over-fitting, and nor too long to lead algorithms to ignore ecological condition changes. Besides, long gaps are considered as one of the primary uncertainty sources of CO_2 flux in the FLUXNET (Aubinet et al., 2012a). As a result, studying the effects of the gap lengths, as well as the window length whereby an algorithm is trained are both critical challenges associated with the environmental data gap-filling.

Apart from the limitations and disadvantages of the mentioned algorithms, gap-filling of fluxes (i.e. NEE) experiences some other challenges that make it necessary to find or develop new gap-filling algorithms. That is because the current methods are not flexible enough to perform well in special occasions or extreme values (Kunwor et al., 2017), and there is almost no room to optimise them to improve their outcome (Moffat et al., 2007). Moreover, even using the best available algorithm, such as ANNs, the model (gap-filling) uncertainty still accounts for a sizable proportion of the total uncertainties, especially when the gaps are relatively long. Since the 2000s when MDS and ANNs were chosen as the most reliable gap-filling methods for EC flux observations, many new ML and

Formatted: Indent: First line: 0 cm, Space After: 0 pt, Line spacing: single optimisation algorithms have been developed and used in varieties of scientific fields. Some of which have shown superiority over ANNs, either individually or as a part of a hybrid or ensemble model, e.g. (Gani et al., 2016). As a result, comparing the cutting-edge algorithms with the current standard ones can show whether there is any room to improve the gap-filling process within the field. According to the concerns mentioned above, this paper had two objectives. a) To find out the impact of different window lengths on the performance of each algorithm. And b) to evaluate the performance of traditional and new gap-filling techniques, separately for fluxes and their meteorological drivers, particularly soil moisture, for this has always been a challenging variable to gap-fill for a couple of reasons, such as of the biology and heterogeneity of soil parameters. To address these objectives, we utilised nine different algorithms (Extreme Gradient Boost (XGB), Random Forest Algorithm (RF), Artificial Neural Networks (ANNs), Marginal Distribution Sampling (MDS), Classic Linear Regression (CLR), Support Vector Regression (SVR), Elastic net regularisation (ELN), Panel Data (PD) and Prophet Forecast Model (FBP)) to fill the gaps of the major fluxes, and eight of them (excluding MDS) to fill the gaps of the environmental drivers. We then assessed their relative performance to evaluate potentially better ways to fill EC flux data. To test the approaches, we used five flux towers from the OzFlux network. To evaluate the performance of these algorithms, nine scenarios for gaps were planned – from a day to a whole year - and applied to the datasets, and different common performance metrics (e.g. RMSE, MBE, etc.), as well as visual graphs were used.

Deleted: window lengths

Deleted: for the OzFlux Network

Deleted: To

Deleted: eight

Formatted: Font: 10.5 pt, Not Bold, Complex Script Font: +Headings CS (Times New Roman), 10.5 pt, Not Bold, (Complex) Persian (Iran)

Deleted: and the major fluxes

2. Materials and methods

To address the first objective of this research, nine different gap lengths were superimposed to the datasets, i.e. 1, 5, 10, 20, 30, 60, 90, 180 and 365 days. To address the second objective, we chose nine different algorithms to fill the gaps, including a wide variety of different approaches, e.g. from a simple algorithm like CLR to cutting-edge ML algorithms, such as XGB (MDS was not used for the environmental drivers). The data used in this paper came from five EC towers of the OzFlux Network, i.e. Alice Springs Mulga, Calperum, Gingin, Howard Springs and Tumbarumba from 2012 to 2013, with a time resolution of 30 minutes, except for Tumbarumba (60 minutes). Additionally, data coming from three additional sources outside of the network were also used as ancillary data to help the algorithms fill the gaps of environmental drivers.

Deleted: data of

Deleted: window

Deleted: considered to train and test the algorithms

Deleted: eight

Deleted: form

Deleted: 2011

2.1. Data

The data used for this research came from OzFlux, which is the regional Australian and New Zealand flux tower network that aims to provide a continental-scale national research facility to monitor and assess Australia's terrestrial biosphere and climate (Beringer et al., 2016a). As described in (Isaac et al., 2017), all OzFlux towers continuously measure and record 28 environmental features at resolutions up to 10 Hz, and use a 30 min averaging period, with a few exceptions (data are available from (http://data.ozflux.org.au/portal). Besides, the network acquires additional data from the Australian Bureau of Meteorology (BoM), the European Centre for Medium-Range Weather Forecasting (ECMWF), and the Moderate Resolution Imaging Spectroradiometer (MODIS) on the TERRA and AQUA satellites (Isaac et al., 2017). These additional data, also known as ancillary data, provide alternative data for gap-filling flux tower datasets (Isaac et al., 2017). As explained in (Isaac et al., 2017), OzFlux uses the BoM automated weather station (AWS) datasets to gap-fill the meteorological data, the BoM weather forecasting model (ACCESS-R) for radiation and soil data from 2011 onward, and MODIS MOD13Q1 for Normalised Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI). Moreover, the data provided by BIOS2, a physically-based model-data integration environment for tracking Australian carbon and water (Haverd et al., 2015), were also used as another ancillary source for varieties of environmental features. Current ACCESS-R and MODIS data are available from the BoM OPeNDAP (http://www.opendap.org/) server and TERN-AusCover data (http://www.auscover.org.au/), respectively.

The datasets were used in this research came from five towers amongst the OzFlux Network between 2012 and 2013, each representative of a different climate and land cover of Australian ecological conditions; i.e. Alice Springs Mulga: Tropical and Subtropical Desert, Calperum: steppe, Gingin: Mediterranean, Howard Springs: Tropical Savanna, Tumbarumba: Oceanic (Table 1) (Beringer et al. 2016). The datasets included 15 meteorological drivers as well as three major fluxes recorded (Table 2) based upon EC technique at a 30-minute temporal resolution, except for Tumbarumba, which was hourly. Additionally, relevant ancillary datasets for the mentioned towers were used to follow the OzFlux Network gap-filling protocol. Each dataset was quality checked at three levels based on the OzFlux Network protocol described in (Isaac et al., 2017) and applied using PyFluxPro ver. 0.9.2. To address the underestimation of canopy respiration by EC measurements at night, we used the CPD method of (Barr et al., 2013) to reject nightly records when the friction velocity fell below the threshold value of each site. After dismissing the inappropriate measurements, overall coverage of 72-88 % and 21-48 % were achieved for diurnal and nocturnal records during 2013 (the year to which the artificial gaps were superimposed), respectively.

Deleted: 2011

Formatted: Font: 10.5 pt, Complex Script Font: 10.5 pt

Deleted: Table 1

Formatted: Font: 10.5 pt, Complex Script Font: 10.5 pt

Deleted: Table 2

Table 1. The information of the five towers that their data were used, including their name, location, dominant species and climate.

Site	Location	Species	Climate	Latitude, Longitude (degree)
Alice Springs Mulga [AU-ASM]	Pine Hill cattle station, near Alice Springs, Northern Territory	Semi-arid mulga (Acacia aneura) ecosystem	Tropical and Subtropical Desert Climate (Bwh)	-22.2828° N, 133.2493° E
Calperum [AU-Cpr]	Calperum Station, 25 km NW of Renmark, South Australia	Recovering Mallee woodland	Steppe Climate (Bsk)	-34.0027° N, 140.5877° E
Gingin [AU-Gin]	Swan Coastal Plain 70 km north of Perth, Western Australia	Coastal heath Banksia woodland	Mediterranean Climate (Csa)	-31.3764° N, 115.7139° E
Howard Springs [AU-How]	E of Darwin, NT	Tropical savanna (wet)	Tropical Savanna Climate (Aw)	-12.4943° N, 131.1523° E
Tumbarumba [AU-Tum]	Near Tumbarumba, NSW	Wet temperate sclerophyll eucalypt	Oceanic climate (Cfb)	-35.6566° N, 148.1517° E

Table 2. List of variables and their units used in this research, including the three main fluxes and their environmental drivers.

List of variables	Units
Drivers:	
Ah	Absolute Humidity (g m^{-3})
Fa	Available energy (W m^{-2})
Fg	Ground heat flux (W m^{-2})
Fld	Downwelling long-wave radiation (W m^{-2})
Flu	Upwelling long-wave radiation (W m^{-2})
Fn	Net radiation (W m^{-2})
Fsd	Downwelling short-wave radiation (W m^{-2})
Fsu	Upwelling short-wave radiation (W m^{-2})
ps	Surface pressure (kPa)
Sws	Soil water content (m m^{-1})
Ta	Air temperature (C)
Ts	Soil temperature (C)
Ws	Wind speed (m s^{-1})
Wd	Wind direction (deg)
Precip	Precipitation (mm)
q	Specific Humidity (kg kg^{-1})
Fluxes:	
Fc (also NEE)	CO_2 flux (μmol m^{-2} s^{-1})
Fh (also H)	Sensible heat flux (W m^{-2})
Fe (also LE)	Latent heat flux (W m^{-2})

The datasets whereby each environmental variable was gap-filled are shown in Table 3. For each of these variables, the same variable of the ancillary source was used to fill the gaps. For instance, to gap-fill Ah, the Ah records of AWS, ACCESS-R and BIOS2 were used. To gap-fill the missing values of fluxes, i.e. Fc (NEE), Fh (H) and Fe (LE), eight drivers were used as follows: Ta, Ws, Sws, Fg, VPD, Fn, q and Ts based on a combination of RF feature selection and testing out a series of feature combinations. Different libraries of Python Programming Language (ver. 3.6.4) were utilised for training and testing the algorithms, i.e. xgboost for XGB, fbprophet for FBP, statsmodels for PD and sklearn for the rest of algorithms. Each algorithm was tuned up individually using grid search, and the number of nodes, layers, irritations, etc. were chosen therefor.

Formatted: Font: 10.5 pt, Complex Script Font: 10.5 pt

Deleted: Table 3

Deleted: trial and error

Deleted: ir

Table 3. The ancillary sources whereby each environmental driver was gap-filled.

List of variables (y)	Ancillary Source
Drivers:	
Ah	AWS, ACCESS-R, BIOS2
Fa	ACCESS-R, BIOS2
Fg	ACCESS-R, BIOS2
Fld	ACCESS-R, BIOS2
Flu	ACCESS-R, BIOS2
Fn	ACCESS-R, BIOS2
Fsd	ACCESS-R, BIOS2
Fsu	ACCESS-R, BIOS2
ps	AWS, ACCESS-R

Sws	ACCESS-R, BIOS2
Ta	AWS, ACCESS-R, BIOS2
Ts	ACCESS-R, BIOS2
Ws	AWS, ACCESS-R
Wd	AWS, ACCESS-R
Precip	AWS, ACCESS-R, BIOS2

2.2. Gap-filling algorithms

Eight imputation algorithms for estimating 15 environmental drivers and 9 algorithms for the 3 major fluxes were picked out to make the comparison. These algorithms were used in a way that a variety of approaches were tested, from the standard methods like ANNs and MDS, to the newer algorithms which rarely or never been used in the field, such as Extreme Gradient Boosting and panel data.

Marginal Distribution Sampling (MDS)

As introduced by Reichstein (Reichstein et al., 2005), the MDS is an enhanced Look-up Tables method, which considers both the covariation of fluxes with meteorological variables and the temporal auto-correlation of the fluxes (Aubinet et al., 2012b). Alongside the ANNs, the MDS is considered as one of the standard gap-filling methods for flux data amongst the FLUXNET, and is selected in this study to help the community to have a clear idea of the performance of other algorithms. Unlike the other algorithms used in this research, we used Fsd, Ta and VPD as the input features for the MDS. The PyFluxPro ver. 0.9.2 was used to apply the algorithm (modified code used for the gaps longer than 10 days).

Artificial Neural Networks (ANN)

Rooted in the 1950s, artificial neural networks are ML methods inspired by biological neural networks and are classified as supervised learning methods (Dreyfus, 1990; Farley and Clark, 1954). ANN work based on several connected units called nodes, which are used to mimic the functionality of a neuron in an animal brain by sending and receiving signals to other nodes. The ANN technique used in this paper was Multi-layer Perceptron regressor, which optimises the squared-loss using stochastic gradient descent. Sklearn.neural_network.MLPRegressor was used to apply this method in Python, and its hyperparameters were 800 and 500 for "hidden_layer_sizes" and "max_iter", respectively based on grid search. ANN are one of the current standard approaches for gap-filling in FLUXNET and in this research were picked out as a performance reference for other algorithms.

Classical Linear Regression (CLR)

Deleted: Eight

Deleted: a

Deleted: According to the literature, since the performance of ANNs had been either equal or slightly better than MDS (Kim et al., 2020; Moffat et al., 2007), we did not use MDS.

Formatted: Justify Low

A classical linear regression is an equation developed to estimate the value of the dependent variable (y) based on independent values (x_i). In contrast, each x_i has its specific coefficient and an overall intercept value. In this method, these coefficients are determined by minimising the squared residuals (errors) of estimated vs observed values, called least squares. A CLR algorithm can be formulated as follows (Freedman, 2009):

$$y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_i X_i + \varepsilon \quad (1)$$

where y is the dependent variable, α is the interception, Xs are independent variables, and βi is coefficient of Xi, and ε is the error term. We chose this algorithm as a baseline to find out how better more complicated algorithms can estimate dependent variables comparatively.

Random Forests (RF)

Random forest, a supervised ML algorithm, used for both classification and regression, consists of multiple trees constructed systematically by pseudorandomly selecting subsets of components of the feature vector, that is, trees constructed in randomly chosen subspaces (Ho, 1998). RF algorithm has been developed to control the overcome over-fitting problem, a commonplace limitation of its preceding decision tree-based methods (Ho, 1995, 1998). Sklearn.ensemble.RandomForestRegressor was used to apply this method in Python, and the hyperparameters used were 5 and 1000 for “max_depth” and “n_estimators”, respectively based on grid search.

Support Vector Regression (SVR)

As a non-linear method, support vector regression was developed based on Vanpik’s concept of support vectors theory (Drucker et al., 1997). An SVR algorithm is trained by trying to solve the following problem:

minimise $\frac{1}{2}\,\|w\|^2$

subject to $\begin{pmatrix} y_i - \langle w, x_i \rangle - b \le \varepsilon, \\ \langle w, x_i \rangle + b - y_i \le \varepsilon, \end{pmatrix}$

where x_i and y_i are training sample and target value in a row. The inner product plus intercept $\langle w, x_i \rangle + b$ is the prediction for that sample, and ε is a free parameter that serves as a threshold. sklearn.svm.SVR was used to apply this method in Python, and the hyperparameters that used were 1 and 0.001 for “C” and “gamma”, respectively based on grid search.

Elastic net regularisation (ELN)

The elastic net is a linear regularised regression method that exerts small amounts of bias by adding two penalty components to the regressed line to decline the coefficients of independent variables and thus, provides better long-term predictions. Given that these two penalty components come from ridge regression and LASSO, the elastic net is considered as a hybrid model consists of ridge and LASSO regressions, overcoming the limitations of both. The estimates from the ELN method can be formulated as below (Zou and Hastie, 2005):

$$\hat{\beta}(elastic\ net) = \frac{\left(|\hat{\beta}(OLS)| - {}^{\lambda_1}/_{2}\right)}{1 + \lambda_2} sgn\{\hat{\beta}(OLS)\} \tag{2}$$

where $\hat{\beta}$ is the coefficient of each ELN independent variable, λ_1 and λ_2 are penalty coefficients of LASSO and ridge regression respectively, $\hat{\beta}(OLS)$ is the coefficient of an independent variable calculated based on ordinary least squares, and *sgn* stands for the sign function:

$$sgn(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases} \tag{3}$$

The ELN regression is good at addressing situations when the training datasets have small samples or when there are correlations between parameters. sklearn.linear_model.ElasticNet was used to apply this method in Python, and the hyperparameters used were as follows: {'alpha': 0.01, 'fit_intercept': True, 'max_iter': 5000, 'normalize': False} based on grid search.

Panel data (PD)

Panel data is a multidimensional statistical method, mainly used in econometrics to analyse datasets, which involve time series of observations amongst individual cross-sections (Baltagi, 1995) usually based on ordinary least squares (OLS) or generalised least squares (GLS). A two-way panel data model consists of two extra components above a CLR as follows (Baltagi, 1995; Hsiao et al., 2002; Wooldridge, 2008):

$$y_{it} = \alpha + \beta X_{it} + u_{it} \qquad i = 1, 2, \ldots, N;\ \ t = 1, 2, \ldots, T \tag{4}$$

$$y_{it} = \alpha + \beta X_{it} + \mu_i + \lambda_t \tag{5}$$

where i and t denote the cross-section and time series dimension in a row, y is a dependent-variable vector, X is an independent variable matrix, α is a scalar, β is the coefficient of the independent-variable matrix, μ_i is the unobservable individual-specific effect, and λ_t is the unobservable time-specific effect. Panel data abilities to provide a holistic analysis of different individuals, as well as determining the specific impact of every single time caused its superiority over CLR. Since PD requires cross-sections to be applied, we used a cross-section tower for each of the main five tower as follows: Ti Tree East for Alice Springs Mulga, Whroo for Calperum, Great Western Woodlands for Gingin, Daly River for Howard Springs, and Cumberland Plain for Tumbarumba. The cross-section towers were chosen based on their distances (the closest ones with common years of data).

Formatted: Font: Palatino Linotype, 10.5 pt, Complex Script Font: 10.5 pt

Formatted: Font: Palatino Linotype, 10.5 pt, Complex Script Font: 10.5 pt

Formatted: Font: Palatino Linotype, 10.5 pt, Complex Script Font: 10.5 pt

Extreme Gradient Boost (XGB)

Extreme gradient boost is a reinforced method of Gradient Boost introduced in 1999 that works based on parallel boosted decision trees and similar to RF can be used for a variety of data processing purposes including classification and regression (Friedman, 2002; Jerome H. Friedman, 2001; Ye et al., 2009). XGB method is resistive to over-fitting and provides a robust, portable and scalable algorithm for large-scale boosting decision-trees-based techniques. sklearn.ensemble.GradientBoostingRegressor was used to apply this method in Python, and its hyperparameters were chosen based on grid search as follows: {'learning_rate': 0.001, 'max_depth': 8, 'reg_alpha': 0.1, 'subsample': 0.5}.

Formatted: Indent: First line: 1.27 cm

The Prophet Forecasting Model (FBP)

The Prophet Forecasting Model, also known as “prophet”, is a time series forecasting model developed by Facebook to manage the common features of business time series and designed to have intuitive parameters that can be adjusted without knowing the details of underlying model (Taylor and Letham, 2017). A decomposable time series model was used (Harvey and Peters, 1990) to develop this model, with three main components: trend, seasonality, and holidays as the equation below (Taylor and Letham, 2018):

$$y(t) = g(t) + s(t) + h(t) \tag{6}$$

where g(t) is the trend function, which models non-periodic changes, s(t) is a function to represent periodic changes, e.g. seasonality, and h(t) assesses the effects of potential anomalies which occur over one or more days, e.g. holidays.

Deleted: mom

2.3. The gap scenarios

In order to find out the effect of gap size on the performance of our gap-filling algorithms, the data of nine different gap windows (i.e. 1, 5, 10, 20, 30, 60, 90, 180 and 365 consecutive days) were removed randomly from the datasets during 2013. Afterwards, the data from 2012 to 2013 were used to train the algorithms. Finally, the trained algorithms were used to fill the artificial gaps superimposed to the datasets. The entire process permutated five times in each scenario to ensure the performance was not sensitive to the gap period. As such, 15 variables, 9 window lengths, 8 gap-filling methods (MDS excluded), and 5 permutations across 5 towers resulted in 27000 computations for the meteorological features. Similarly, 3 fluxes, 9 window lengths, 9 gap-filling methods, and 5 permutations across 5 towers resulted in 6075 computations for the major fluxes, overall.

2.4. Statistical performance measures

Different statistical metrics were used to evaluate the performance of algorithms and enable comparison between measured values from the flux towers with each gap-filling algorithm prediction. These metrics included the coefficient of determination (R-squared) to measure the square of the coefficient of multiple correlations (Devore, 1991), the variance of measured and modelled values (S^2) to indicate how well algorithms could follow the variations of the recorded data, the root mean square error (RMSE), the mean bias error (MBE) to capture distribution and bias of residuals, variance ratio (VR) to compare the variance of estimated values with those of measured, and the Index of Agreement to compare the sum of the squared error to the potential error (Bennett et al., 2013). Abbreviations and formulas of these metrics are illustrated as follows (Bennett et al., 2013):

$$R^2 = \frac{[\sum(p_i - \bar{p})(o_i - \bar{o})]^2}{\sum(p_i - \bar{p})^2 \sum(o_i - \bar{o})^2} \quad (7)$$

$$S^2 = \frac{\sum(x_i - \bar{x})}{N - 1} \quad (8)$$

$$RMSE = \sqrt{\frac{\sum(p_i - o_i)^2}{N - 1}} \quad (9)$$

$$MBE = \frac{\sum o_i - p_i}{N - 1} \quad (10)$$

$$VR = \frac{\sigma_p^2}{\sigma_o^2} \quad (11)$$

$$IoAd = 1 - \frac{\sum_{i=1}^{n}(o_i - p_i)^2}{\sum_{i=1}^{n}(|p_i - \bar{o}| + |o_i - \bar{o}|)^2} \quad (12)$$

where o_i and p_i are individual measured and predicted values respectively, $\bar{o}$ and $\bar{p}$ are the means of o and p, and σ^2 is the variance. S^2 is calculated separately for the observed and predicted values with the respective values defined as x that represents every observed or predicted value. All of these metrics were calculated for each of the gap scenarios, and then the results of different windows were concatenated. Afterwards, the yearly metrics were calculated to avoid Simpson's paradox or any relevant averaging issue as described by (Kock and Gaskins, 2016). Moreover, the average of daily and seasonal differences between the estimated and measured values, as well as the associated variance were calculated and plotted.

Deleted: To find out the effect of gap size on the performance of our gap-filling algorithms, we trained each of them using nine different window lengths (i.e. 1, 5, 10, 20, 30, 60, 90, 180 and 365 days). The gap size for each trained algorithm was chosen as the same size of the corresponding training window, e.g. the gap size for a 20-day training window was 20 days and so on. As such, in every scenario, the entire data of 2013 were used step by step to test the performance of the algorithms as follows: at the first step of each scenario, the gap began from 1 Jan 2013, while its corresponding training window was the same size but came from the preceding period. For instance, for a 30-day gap, the first step included training an algorithm based on the data of Dec 2012 and the testing period of the first month of 2013. In the second step, the data of the first month of 2013 were used for training, while the data of the second month of 2013 was considered as a gap, and this went to the end of 2013 consecutively. As such, for the last step, the training window was the second last 30 days of 2013, and its corresponding gap was the last 30 days of 2013. The only exception of the mentioned training strategy was FBP as it needed a training dataset with at least a year to be developed. Therefore, here, the training data for each gap was all data prior to that gap since the beginning of 2011. Overall, 18 variables, nine window lengths and eight gap-filling methods across five flux towers resulted in 6480 computations.¶

3. Results

3.1. Fluxes

3.1.1 Fc

Even though factors such as Fg and Fn are fluxes, we dealt with them as environmental drivers since they drive the three major fluxes. The metrics used to evaluate the performance of the algorithms (RMSE, R^2, MBE, IoAd and VR) (Table 4) illustrated that overall, the performance of these algorithms, particularly the ML ones, was similar, closely followed by the MDS. The XGB provided the lowest values of RMSE and one of the highest R^2, while the FBP and ELN had the lowest and highest values of RMSE and R^2, respectively. The algorithms, however, showed different levels of sensitivity to the gap lengths, e.g. the CLR and PD showed smaller sensitivity, while the FBP showed the most sensitivity (Figure 1).

Table 4. The average amounts of performance metrics for each gap-filling algorithm regarding Fc, which includes all window lengths and sites, ranked by RMSE using the Tukey's HSD test at the level of 5 per cent.

Algorithm	Mean RMSE	Mean R^2	Mean MBE	Mean IoAd	Mean VR
XGB	3.07 [a]	0.59	-0.43	0.90	0.66
RF	3.12 [a]	0.58	-0.37	0.91	0.71
ANNs	3.13 [a]	0.56	-0.33	0.90	0.69
SVR	3.34 [b]	0.47	-0.32	0.86	0.75
MDS	3.35 [b]	0.51	-0.41	0.85	0.70
PD	3.41 [b,c]	0.48	-0.35	0.81	0.54
CLR	3.44 [b,c]	0.49	-0.36	0.81	0.55
ELN	4.52 [c]	0.43	-0.37	0.73	0.39
FBP	4.15 [d]	0.47	-0.06	0.77	0.68

These outcomes were expected for the XGB as it uses a more regularised model formalisation to control over-fitting (Chen and Guestrin, 2016) which, on paper, leads to better performance as against its ML rivals. The relatively poor performance of FBP was also foreseen for unlike other algorithms, FBP did not use any feature to estimate flux values, other than the previous time series of flux values. However, the weaker performance of the ELN compared to CLR was unforeseen due to by adding two penalty components to the regressed line, and the ELN is supposed to improve the long term prediction compared to the traditional linear regression methods. Tukey's HSD (honestly significant difference) test at the level of five per cent was applied to the results to find out whether the difference amongst the algorithms was significant (Table 4). Where the null hypothesis was there is no significant difference between the mean values of the RMSE. According to the results, there were significant differences between certain algorithms, and the XGB, RF and ANNs were different from the rest, showing that these three performed considerably better. Tukey's HSD test, however, did not reject the second error probability between RF, XGB and ANNs meaning that the three algorithms were not significantly different from each other. This result agrees with the results of (Falge et al., 2001) and (Moffat et al., 2007) in the sense that ANNs are one of the best available gap-filling algorithms, and there is no significant difference amongst the appropriate algorithms. However, the test showed that the performance of the MDS had a significant difference from the ANNs. Finally, it is worth

Formatted
Deleted: Table 4
Moved (insertion) [1]
Deleted: training/testing window…he gap lengths, e.
Formatted
Moved up [1]: The XGB provided the lowest values of
Deleted: 53
Deleted: 56
Deleted: 44
Deleted: 89
Deleted: 59
Deleted: 56
Deleted: 54
Deleted: 38
Deleted: 90
Deleted: 70
Deleted: 57
Deleted: 52
Deleted: 34
Deleted: 89
Deleted: 68
Deleted: 81
Deleted: 33
Deleted: 79
Deleted: 89
Deleted: 45
Deleted: 36
Deleted: 80
Deleted: 53
Deleted: 92
Deleted: 46
Deleted: 37
Deleted: 80
Deleted: 54
Deleted: 01
Deleted: 40
Deleted: 38
Deleted: 72
Deleted: 37
Deleted: 44
Deleted: that
Formatted
Deleted: Table 4…). Where the null hypothesis was

mentioning that Tukey's HSD is well known as a conservative test. That being said, despite no meaningful difference based on Tukey's HSD, XGB and RF might have performed better than ANNs, as the superiority of RF in gap-filling of methane flux over the ANNs, SVR, and MDS has recently been claimed by (Kim et al., 2020).

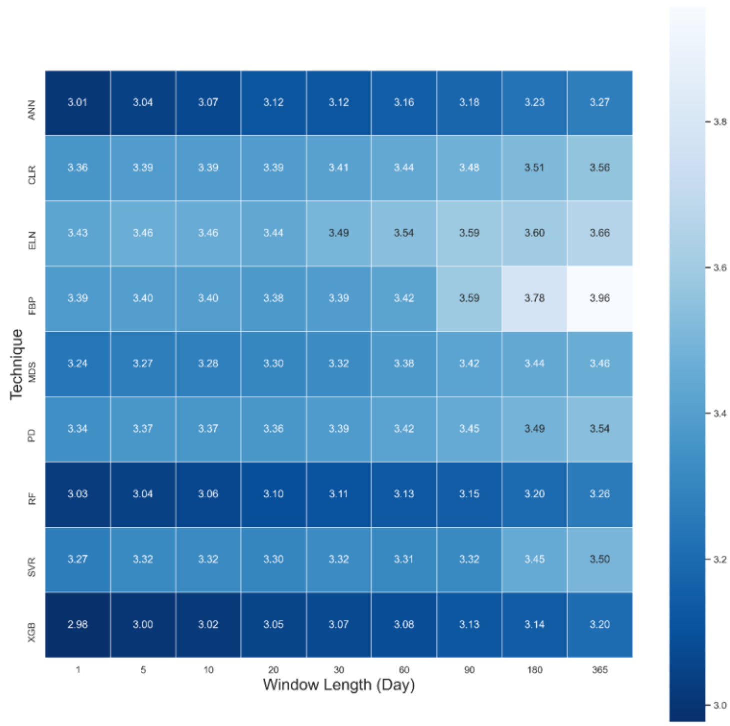


Figure 1. A heat map of mean RMSE values of Fc across all sites based on 9 algorithms and 9 window lengths in 2013.

To address the first objective of this paper, finding out the sensitivity of the gap-filing algorithms to the gap window length, we used the averaged RMSE, R^2 and MBE for each gap size, using the output of all algorithms for all sites (Table 5). The outcome illustrates that the longer the

Formatted: Keep with next

Formatted: Caption

Deleted: *<object>*

Deleted: ¶
Figure 1. A heat map of mean RMSE values of Fc across all sites based on 8 algorithms and 9 window lengths in 2013.¶
¶

Deleted: s

Deleted: each

Deleted: training and testing

Deleted: window length and

Deleted: Table 5

window length got, the bigger the amounts of RMSE became. Yet, no such pattern was recognisable for the R^2 and MBE. As a result, generally, any consecutive gaps longer than 30 days seem to decline the performance of the algorithms noticeably. The phenomenon can be justified by the idea that longer windows do not let the algorithms to accommodate seasonal changes and therefore, different physiological behaviour of the canopy.

Table 5. The average amounts of RMSE, R^2, and MBE for Fc gap-filling based on the window length including the outcome of all sites; the differences of RMSE values were tested using the Tukey's HSD test at the level of 5 per cent.

Window length	Mean RMSE	Mean R^2	Mean MBE
1-day	3.23 [a]	0.53	-0.27
5-days	3.25 [a]	0.52	-0.31
10-days	3.26 [a]	0.51	-0.29
20-days	3.27 [a]	0.51	-0.31
30-days	3.29 [a]	0.51	-0.31
60-days	3.32 [a]	0.49	-0.35
90-days	3.37 [a]	0.51	-0.38
180-days	3.43 [a]	0.50	-0.41
365-days	3.49 [a]	0.49	-0.37

According to the MBE values (Table 4), mainly, all algorithms had negative amounts of MBE, showing overestimation of the Fc values. This bias varied from tower to tower and depended on the window lengths. For instance, absolute amounts of the MBE were bigger in Gingin and Tumbarumba, while considerably smaller (closer to zero) at AliceSprings Mulga and Calperum (Supplementary). The lower leaf area index of the two later sites, and thus their smaller amounts of photosynthesis is likely to be the reason that justifies the outcome. FBP, nonetheless, provided substantially lower mean bias (-0.06) compared to the other algorithms, which varied between -0.32 and -0.43.

Observations from the EC technique often include extremely low or high values, especially at night, when some of the theoretical assumptions might be violated. The nature of the EC technique associated with its practical challenges, often makes it difficult to distinguish between the good data and the noise (Aubinet et al., 2012a; Burba and Anderson, 2010). This problem seems to affect the outcomes of the gap-filling algorithms in this research, as none of them performed ideally in capturing the observed variance (). Even though RMSE, R^2 and IoAd showed the superiority of the XGB, RF and ANNs, the variance ratio between the estimated and measured values revealed different information (Table 4), which is slightly recognisable in Figure 2. The variance ratios (VR) showed that SVR captured the extreme values of Fc better than the other algorithms, 0.75 on average. The other ML algorithms –

Deleted: , particularly for the window lengths equal t ...
Deleted: based on our scenarios (using the same leng ...
Deleted: 60
Deleted: , i.e. 90, 180 and 365 days,
Deleted: made
Deleted: worst
Deleted: 74
Deleted: 49
Deleted: 6
Deleted: 77
Deleted: 49
Deleted: 77
Deleted: 49
Deleted: 79
Deleted: 48
Deleted: 80
Deleted: 48
Deleted: 79
Deleted: 48
Deleted: 81
Deleted: 48
Deleted: 39
Deleted: 88
Deleted: 47
Deleted: 90
Deleted: 46
Deleted: Besides, the metrics of the top three ...
Deleted: results not shown
Formatted: Highlight
Deleted: 33
Deleted: 44
Field Code Changed
Formatted: Font: 10.5 pt, Complex Script Font: 10.5 pt
Deleted: Table 4
Deleted: also
Deleted: Fig. 2
Deleted: 79

plus the MDS- though, performed closely with regard to capturing the extremes that matches both the expectations, and the performance metrics Table 4.

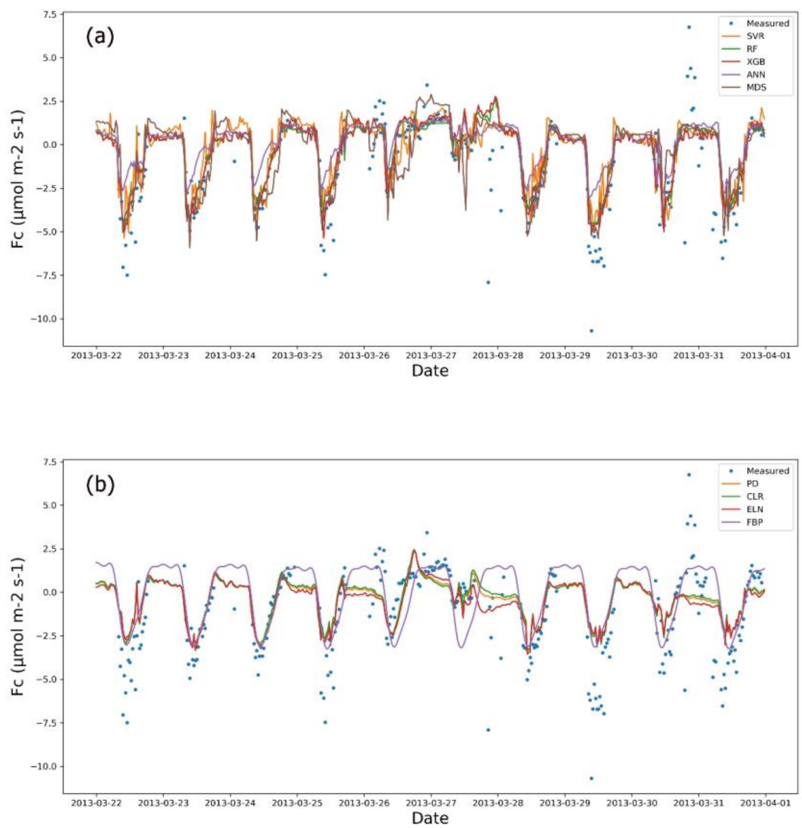
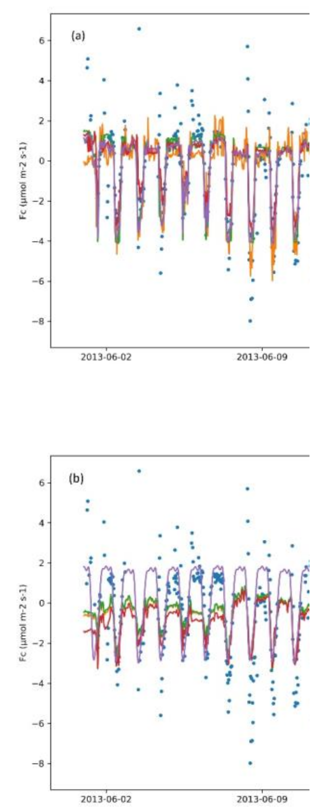


Figure 2. Measured vs estimated values of Fc for Calperum based on a 10-day gap window (March 22 - March 31, 2013).

The linear algorithms, CLR, PD, and ELN, performed worse with respect to the VR compared to the ML algorithms. The estimated versus measured values of Fc for Calperum () confirms the information achieved by the VR. Based on the figure, the ELN, as expected, performed the worst in capturing the fluctuations of Fc (VR = 0.39), while the performance of the other algorithms –apart from the top five- was not considerably better, with the exception of FBP. It is noteworthy that CLR, PD, and ELN frequently predicted nocturnal photosynthesis. Overall, the results showed a significant difference between the top five algorithms (XGB, RF, ANNs, SVR, and MDS) and the others, particularly in capturing the fluctuations and the min-max values of Fc. However, a comprehensive comparision shows a slight superiority of the XGB and RF.

Deleted: XGB, on the other hand, provided smaller VR (0.59) compared with those of the RF (0.70) and ANNs (0.68), especially for the window lengths longer than 10 days (not shown).



Deleted:

Formatted: Caption, Don't keep with next

Deleted: *Figure 2. Measured vs estimated values of Fc for Calperum based on the 30-day window during June (Austral winter) 2013¶*

Deleted: during June 2013

Field Code Changed

Deleted: and VR of 0.79, the SVR captured the extreme values of Fc the best, whereas

Deleted: 37

Deleted: Although the XGB (VR of 0.59) provided relatively well while estimating the maximum values (respiration), it was not capable of assessing the minimum values, thereby provided a constant overestimation of NEE during the day. The RF (VR of

3.1.2 Fe

The performance of algorithms for Fe was similar to that for Fc regarding RMSE, MBE and R^2, as shown in *Table 6*. This similarity was not surprising since these processes are partially coupled via stomatal conductance (Scanlon and Kustas, 2010; Scanlon and Sahu, 2008). Again, the top three ML algorithms performed better, with a significant superiority of the XGB and RF, as shown by the Tukey's HSD (*Table 6*), followed by the ANNs and MDS. Besides, the null hypothesis was not rejected while comparing FBP and SVR, whereas the better performance of the other algorithms was confirmed. As a result, the FBP and SVR provided the most unsatisfactory results in estimating Fe, according to the average values of the RMSE. No significant improvement in RMSE occurred when the gap lengths became shorter than 60 days, meaning that the performance of the algorithms did not vary considerably from a 30-day to a one-day window, especially for the top algorithms (XGB, RF, and ANNs). The results of CLR and PD were very similar to those for Fc, showed lower RMSE and higher R^2 values as against ELN, but the ELN led to slight lower MBE. The MBE values also showed moderately high values for the SVR, meaning that there was an absolute bias in its outcome, which might be related to overfitting. The source of the bias shown by the SVR algorithm (Figure 3), was because it could not capture the minimum values appropriately, resulting in a considerable overestimation. A common issue in estimating Fe values, which had affected all algorithms other than the FBP, was not assessing the negative values. In contrast to Fc results, the ANNs did not perform as solid as the XGB and RF, which could be due to not being able to capture the maximum values as satisfying as its rivals were.

Table 6. The average of metrics for Fe gap-filling based on the algorithms, ranked by RMSE using the Tukey's HSD test at the level of 5 per cent.

Algorithm (Fe)	Mean RMSE	Mean R^2	Mean MBE
XGB	34.95 [a]	0.74	-3.48
RF	35.63 [a]	0.74	-3.33
ANNs	37.77 [a,b]	0.67	-3.94
MDS	41.74 [b,c]	0.64	-3.27
PD	43.28 [b,c]	0.64	-6.35
CLR	43.51 [c]	0.64	-6.66
Eln	44.34 [c]	0.59	-5.13
SVR	46.63 [c,d]	0.59	-20.45
FBP	50.53 [d]	0.52	3.01

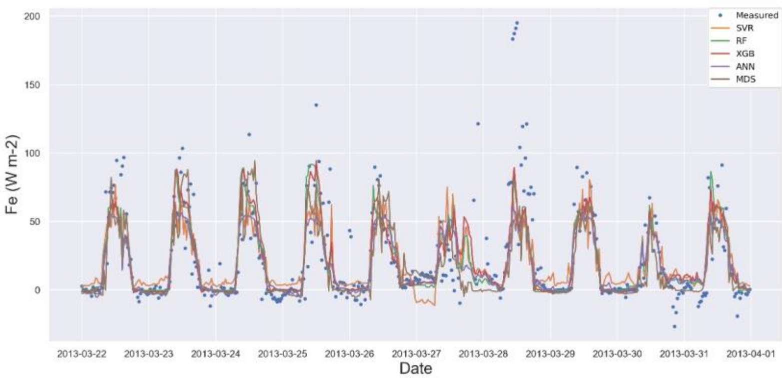


Figure 3. Measured vs estimated values of Fe for Calperum based on a 10-day gap window (March 22 - March 31 2013).

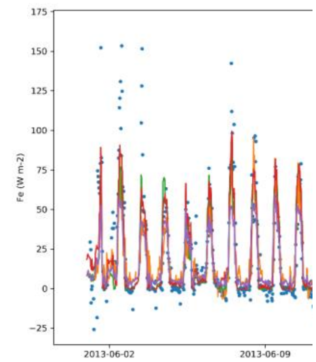
3.1.3 Fh

As with the other flux results, the metrics (RMSE, R^2 and MBE) showed slight superiority of XGB and RF, as well as the inferiority of the SVR and FBP over the other algorithms (Table 7). Likewise, the SVR provided relatively large negative values of MBE, showing considerable overestimation. The Tukey's HSD test of the average RMSE values confirmed that the performance of the FBP was significantly different from the rest at the level of 5 per cent, making FBP the weakest performer for Fh. On the other hand, although there was no significant difference amongst the XGB, RF and ANNs, the first two were considerably superior over the other algorithms as regards the Tukey's HSD test. Like Fe, estimated values of Fh using SVR had a negative bias (Figure 4) because it was not able to provide appropriate estimations of Fh minimum values. In contrast, the ANNs performed the best in capturing the minimum values, while the other top algorithms performed almost equally well. Despite the close performance in capturing the minimum values, ANNs and MDS did not carry out as solid as XGB and RF concerning the overall values, resulted in higher RMSE. Finally, similar to the other fluxes, the PD performed slightly better than the CLR and ELN.

Table 7. The average metrics for Fh gap-filling based on the algorithms, ranked by RMSE using the Tukey's HSD test at the level of 5 per cent.

Algorithm (Fh)	Mean RMSE	Mean R^2	Mean MBE
XGB	37.23 [a]	0.92	-0.21
RF	37.55 [a]	0.91	-0.09
ANNs	40.13 [a,b]	0.90	-0.08
MDS	43.30 [b,c]	0.88	-9.51
SVR	43.80 [b,c]	0.88	0.35
PD	44.96 [c]	0.88	1.36
CLR	45.03 [c]	0.88	1.64
Eln	45.19 [c]	0.87	2.16
FBP	72.91 [d]	0.73	1.07

Deleted:



Formatted: Caption, Don't keep with next

Deleted: *Figure 5. Measured vs estimated values of Fe for*

Formatted: Font: 10.5 pt, Complex Script Font: 10.5 pt

Deleted: Table 7…). Likewise, the SVR provided

Formatted

Deleted: negative …inimum values. In contrast, the

Deleted: 26

Deleted: 93

Deleted: 1.00

Deleted: 38.08…7.55 [a,b]

Deleted: 93

Deleted: 1.35

Deleted: 40.48…0.13 [a,b,c]

Deleted: 2

Deleted: 41

Deleted: 41.83…4.96 [b,]

Deleted: 92

Deleted: -0.27

Deleted: 42.14…5.03 [b,]

Deleted: 92

Deleted: -0.05

Deleted: 42.28…5.19 [b,]

Deleted: 92

Deleted: 0.04

Deleted: SVR

Deleted: 67.19

Deleted: 74

Deleted: 25

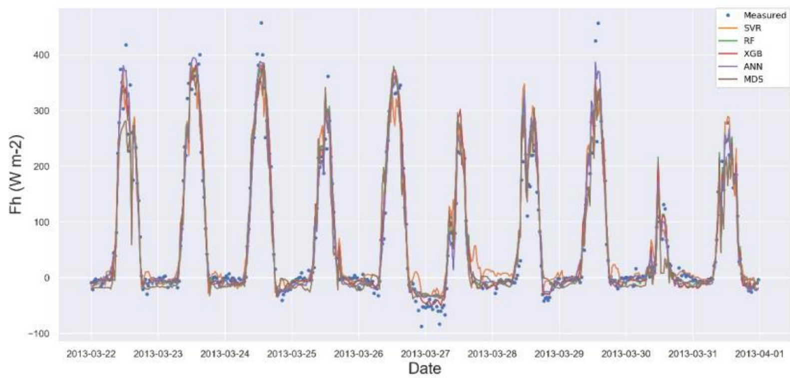
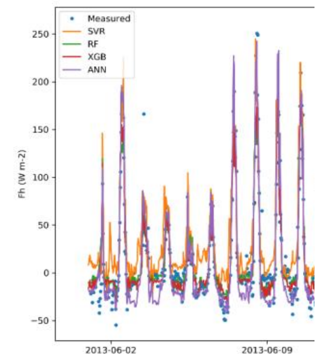


Figure 4. Measured vs estimated values of Fh for Calperum based on a 10-day gap window (March 22 - March 31 2013).

3.2. Meteorological and Environmental Drivers

Since meteorological and environmental drivers are needed to fill the gaps of the three substantial fluxes, Fc, Fe and Fh, the eight algorithms (excluding the MDS) were used to fill the gaps of these drivers. The metrics of R^2, RMSE, and MBE were calculated for all five towers and nine window lengths (16 meteorological and environmental drivers and three fluxes). Overall, for most meteorological drivers, the linear algorithms, especially the CLR and PD, performed slightly better than the ML algorithms such as the XGB, RF, ANNs and SVR, except for Ah, Fg and Fn. This unexpected superiority can be explained based on the two following reasons. Firstly, unlike the fluxes, the input and output features were the same here, e.g. Ta for Ta, which led to strong correlations (e.g. up to 0.99 for atmospheric pressure - ps) as well as strong linear relationships between the independent and dependent features. These strong correlations helped the linear algorithms to perform well, while nullified the ability of the ML algorithms to capture non-linear behaviour of complicated problems. Second, the slight inferiority of ML algorithms could be due to data noise where simple linear algorithms such as the CLR are usually less sensitive to the noise relatively. Therefore, over-fitting is not an issue for them when the number of observations is big enough (i.e. at least 10 to 20 observations per parameter (Harrell, 2014)). The exceptions were Ah, Fn and Fg, for which values were estimated more accurately by the XGB, ANNs and RF, especially the latest one (the RMSE of 28.91 versus 33.92 provided by the RF and CLR for Fg, respectively). Tukey's HSD test for the mean RMSE values of Fg confirmed that The XGB, ANNs and RF provided better results at the level of 5 per cent, while, like all other fluxes and drivers, the FBP confirmed to be the worst algorithm (Table 8). Yet, according to the same test for the other drivers, there was not any significant difference between the algorithms, other than the FBP, which provided the most significant mean values of the RMSE (results not shown). Importantly, though, none of the algorithms offered adequate estimations for soil moisture (Sws), particularly in drier regions. This weak performance happened because Sws



Deleted:

Formatted: Caption, Don't keep with next

Deleted: *Figure 6. Measured vs estimated values of Fh for Calperum based on a 60-day window during June 2013¶*

Deleted: 30.23

Deleted: 35.24

Formatted: Font: 10.5 pt, Complex Script Font: 10.5 pt

Deleted: Table 8

changes dramatically during rainfall in a pulsed manner often from zero to saturation in short space of time, whereas, the algorithms had been trained based on the datasets mostly reflecting non-rainy periods. These datasets, consequently, could not fit the algorithms in a way that they could estimate Sws accurately when precipitation occurs and the soil moisture increases dramatically. For instance, in a wet region like Tumbarumba, where the soil faces rainy days frequently, the time series are much less spikey. Thus, the overall performance was better in these regions compared with the drier ones, e.g. R^2 of 0.45 and 0.26 on average for Tumbarumba and Calperum, respectively. Besides, the dataset used to gap-fill the soil moisture was a model derivation from gridded data or regional reanalysis and therefore, can be not close to reality. Another challenge of estimating soil moisture comes from the low spatial coherence of soil moisture is that it can be extremely different just a couple of hundred metres away, due to storms, topography, soil structure heterogeneity, etc. (Reichle et al., 2004; Sahoo et al., 2008).

Table 8. The average amounts of RMSE for Fg gap-filling based on the algorithms, using the Tukey's HSD test at the level of 5 per cent.

Algorithm (Fg)	Mean RMSE
RF [a]	28.91
XGB [a, b]	29.19
ANNs [b, c]	29.58
SVR [c]	31.46
CLR [d]	33.92
PD [d]	33.93
ELN [d]	34.09
FBP [e]	39.10

4. Discussion

Table 9. The name and the abbreviation of the gap-filling algorithms.

Algorithm abbrevation	Full name
XGB	Extreme Gradient Boost
RF	Random Forest Algorithm
ANNs	Artificial Neural Networks
MDS	Marginal Distribution Sampling
SVR	Support Vector Regressi
CLR	Classical Linear Regression
PD	Panel data
ELN	Elastic net regularisation
FBP	The Prophet Forecasting Model (Facebook Prophet)

All algorithms (Table 9) performed similarly in estimating the meteorological and environmental drivers (turbulent fluxes included) across all stations, except the FBP, which performed poorly for it did not use any ancillary data. The best results were achieved for the 30-day gaps and shorter, while the worst results obtained for the most extended windows, 180 and 365 days. Although

Deleted: 43 …5 and 0.25
Deleted: 30.17
Deleted: 30.70
Deleted: 30.86
Deleted: 32.77
Deleted: 34.93
Deleted: 34.94
Deleted: 94

most of the algorithms performed almost equally well in estimating meteorological and environmental drivers, the linear algorithms, the CLR, ELN and PD, performed slightly better (not significant using a Tukey's HSD test, though). The only clear exception was Fg, for which the RF provided more accurate and robust estimations. The ML algorithms and MDS, on the other hand, showed their superiority over the linear algorithms while estimating the main fluxes, Fc, Fe and Fh. For Fc, the XGB, RF and ANNs performed significantly better than the FBP and all linear algorithms, i.e. the CLR, PD and ELN, yet, followed closely by the SVR and MDS. The superiority of the ML algorithms, as well as their close performance, agreed with the results of previous researches, e.g. (Falge et al., 2001; Moffat et al., 2007), that showed the superiority of non-linear algorithms and no significant difference amongst the top algorithms in estimating Fc. Besides, the slight superiorities of XGB and RF over ANNs, mainly unnoticeable by a conservative test like Tukey's HSD, confirms RF performs better regarding the EC flux gap-filling (Kim et al., 2020).

The XGB was the most novel ML algorithm used in this research and based on the most performance metrics provided comparatively robust results in estimating the fluxes. In estimating the meteorological drivers though, the XGB did not show any superiority over the other algorithms, especially the linear ones. Moreover, the XGB needed four to six times longer time to be trained and tunned, making it a less feasible algorithm when time or the processing power are important factors or several years of data are needed to be gap-filled. Hence, we do not recommend the XGB as an alternative to the current alternative algorithms. Nevertheless, because of its local superiorities, this algorithm might be suitable to use in an ensemble model alongside the algorithms with different weakness points.

The RF was the best all-around algorithm amongst the nine algorithms used in this study, providing the best consistant and robust estimates of the fluxes (similar to XGB) but also being less complicated and performing faster than the XGB. The RF also provided the best results for Fg, where the linear algorithms did not perform well. This superiority of this algorithm over ANNs, MDS, and SVR has been proved by (Kim et al., 2020) for gap-filling of methane, showing that it is worth testing the RF for other towers, and fluxes across the FLUXNET.

The ANNs estimated the fluxes better than the linear algorithms, notably for Fc, yet not as robust as the XGB and RF in general. For Fc and Fh, the ANNs provided bias, mainly due to overestimation of minimum values when the window lengths were longer than 30 days. However, since the superiority of the XGB and RF was not considerable, it is difficult at this point to suggest using XGB or RF as better alternatives. That is because ANNs have been checking out for a long time in different locations and considered as one of the most reliable algorithms in the field for more than a decade (Aubinet et al., 2012a; Hagen et al., 2006; Kunwor et al., 2017; Moffat et al., 2007). In other words, the superiority of RF, needs to happen in several future studies to convince the network to suggest RF instead of ANNs, or identify it as another standard method. Furthermore, there are a wide variety of different ANNs algorithms used in the field (Beringer et al., 2016b; Hagen et al., 2006; Isaac et al., 2017; Kunwor et al., 2017; Moffat et al., 2007), and this minor superiority of RF and XGB cannot be generalised without enough additional proves. As such, we suggest other researches to use the RF, especially regarding Fh and Fc alongside the ANNs to find out which one performs better in the

Deleted: of

Deleted: SVR,

Deleted: However, the XGB failed to capture the minimum values of Fc as against the SVR and RF, and thus, provided biased results, while assessing the maximum values of Fc well.

Deleted: , especially for long gaps

Deleted: is

Deleted: eight

Deleted: captured

Deleted: both minimum and maximum values of Fc

Deleted: Unlike the RF, all other algorithms generally struggled with estimating either minimum or maximum values of major fluxes, comparatively. It

Deleted: Another advantage of the RF over the XGB was that it required less time (approximately a quarter) for training, which was a challenge while using the XGB.

Formatted: Font: Palatino Linotype, 10.5 pt, Font color: Auto, Complex Script Font: 10.5 pt

Formatted: Font: Palatino Linotype, 10.5 pt, Font color: Auto, Complex Script Font: 10.5 pt

Formatted: Font: Palatino Linotype, 10.5 pt, Font color: Auto, Complex Script Font: 10.5 pt challenging scenarios, e.g. when the gaps are long. Another option is to develop ensemble models using since, according to the literature, there is no room to improve the results substantially based on a single algorithm (Moffat et al., 2007). Besides, a model with a higher level of flexibility is required in the field (Hagen et al., 2006; Kunwor et al., 2017; Richardson and Hollinger, 2007). Finally, according to the environmental drivers, The ANNs, like the other ML algorithms, could not show a consistent superiority over the linear algorithms. Therefore, we do not recommend using ML algorithms in such scenarios, except for Fg, for which RF seems to be a better option.

Deleted: On the other hand

The MDS performed close to, yet not as well as the XGB, RF, and ANNS in gap-filling the fluxes. Its performance was close to the SVR, but was more reliable regarding Fe and Fh. It is worth mentioning that this performance was achieved despite the fact that the MDS was using fewer input features. Its performance, however, was comparable with the ML algorithms, particularly when the gap lengths were relatively shorter (smaller than 10 days). As such, we recommend using the MDS when the gaps are not long and/or the available input features are limited, especially considering that the MDS performs significantly faster than the ML algorithms, and is easier to use.

The SVR showed consistent inferiority over the other ML algorithms and did not fulfilled our expectations, neither for the meteorological drivers nor for the major fluxes. The only strength of the SVR was that it captured the extreme values better than any other algorithm. However, according to its larger RMSE amounts, the mentioned advantage seems to be achieved suspiciously and might have occurred due to over-fitting. This dubious performance shows the SVR is more vulnerable to the over-fitting issues regarding these types of data. Hence, we suggest the SVR not to be used in any kind of environmental modelling related to the reviewed drivers and fluxes, whatsoever.

Deleted: , as revealed in the plots of Fc

The CLR, the simplest algorithm used in this research, provided a comparatively acceptable performance in estimating the meteorological drivers, except for Fg. This algorithm, however, could not perform well in assessing the fluxes, especially Fc, mainly because of its inability to capture the extreme values caused by the non-linear nature of Fc. Overall, considering the CLR simplicity, resource-saving and robust performance for drivers, this algorithm seems to be the most suitable way to fill the gaps of meteorological parameters in similar scenarios, where the same ancillary dataset are available.

Deleted: is

The PD performed slightly better than the CLR, yet it could not fulfil the expectations to show a significant superiority over the other linear algorithms used in the research. This unforeseen weak performance can be explained due to a couple of reasons. First, one of the assumptions of using the PD is that the behaviour of the cross-sections, here towers, is similarly under the similar conditions (the independent variables), and the only thing leads to the difference is the specific characteristics of each individual cross-section. Contrariwise, it seems that the five towers selected in this research violated this assumption due to their absolute different ecosystems. Based on the previous studies in which the PD performed satisfying (Izady et al., 2013, 2016; Mahabbati et al., 2017), (Izady et al., 2016) and (Mahabbati et al., 2017), it appears that a decent level of homogeneity is vital for the PD to perform satisfactorily. As in all previous cases, the ecosystem of the cross-sections had significant similarities, and the distance between them were tens to hundreds of kilometres, not thousands. Therefore, the characteristics of cross-sections, such as radiation, climate, rainfall, etc. had considerable more similarity and homogeneity compared with the towers used in this research. Finally, it is worth mentioning that PD has been commonly used to analyse the time series with a time resolution of weekly or longer, with some exceptional daily-scale cases. In this research, the resolution of data was half-hourly instead, which dramatically increased the computational demands of the algorithm, led to days of processing for a single run. This demand happened because the algorithm creates a dummy variable for each time step and the relevant matrix of variables becomes too large to compute by a regular PC. Considering the expenses of this algorithm, we recommend other researches not to use PD when the time resolution is shorter than daily. Despite the limitation, we still encourage further using of PD whenever there is a decent level of homogeneity amongst the cross-sections and the time resolution is daily or longer (ideally weekly or monthly).

The ELN, as a hybrid linear model, did not show any superiority over the CLR, despite its modifications to provide more accurate estimations. Even though ELN performed well in estimating the drivers with slight supremacy in some occasions, e.g. Fld, the CLR is a more proper algorithm to choose for gap-filling the drivers due to its simplicity and less calculation requirement.

The FBP was a unique algorithm used in this research, as it did not use any independent variables to estimate the values of drivers and fluxes. The FBP performance was significantly more unsatisfactory than the other algorithms. Therefore FBP cannot be considered as a reliable alternative for current algorithms to fill the gaps, especially the long ones.

Given that some of the environmental drivers affect the Fc differently during the day versus night, separating the diurnal and nocturnal datasets to train the algorithms possibly entails an improvement in the outcome. Mainly because of the u^* threshold filtering and other problems associated with the nocturnal period, the portion of diurnal data is generally, by far, outweighs the nocturnal data portion, which potentially leads to a bias in the algorithm. The same challenge has associated with soil moisture estimation, as the behaviour of the system on sunny days is utterly different from its conduct during the rainy periods. Moreover, the system memory and the antecedent condition are undeniable features associated with soil moisture (Ogle et al., 2015). Therefore, using the models that are capable of addressing these considerations are more likely to improve the estimations.

Moved (insertion) [2]

Moved (insertion) [3]

Deleted: solution as number 4 is suggested

Deleted: for

5. Conclusions

Eight different gap-filling algorithms for estimating 16 meteorological drivers as well as Nine algorithms for the three key ecosystem turbulent fluxes (sensible heat flux (Fh), latent heat flux (Fe), and net carbon flux (Fc)) were investigated and their performance evaluated based on the datasets of five towers in Australia. Overall, three ML algorithms, XGB, RF and ANNs, performed nearly equally well and significantly better than their linear rivals (the CLR, PD, and ELN) in estimating the flux values. However, the linear algorithms performed almost as equally well as the ML algorithms in assessing the meteorological drivers. Amongst these nine algorithms, the RF and XGB showed the highest level of robustness and reliability in estimating the Fc, Fe, and Fh. The PD was expected to perform better than the linear methods and hoped to compete with the ML algorithms in estimating the fluxes, but it failed to do so. The SVR was the only ML algorithm that did not perform at the same

Deleted: eight

Deleted: , as its closest rival, the XGB, could not capture the minimum values equally well, despite providing slightly better RMSE and R^2

level as the rest ML algorithms and was suspected of enduring over-fitting issues, while the MDS performed somewhere in between. Considering the outcomes of the other researches undertaken in the OzFlux Network, e.g. (Cleverly et al., 2013; Isaac et al., 2017), none of the ML algorithms used in this research was proven to provide substantially better flux estimations compared with the standard method (ANNs). Nonetheless, amongst the algorithms tested in this research, the RF showed some potential capabilities as an alternative due to its more consistent performance regarding the long gaps. Eventually, we recommend suggestions below to improve the results for similar prospective researches, as well as the QC and gap-filling procedure of OzFlux Network:

1) Since the RF remained more consistent compared to its competitors -including the ANNs-, It is a good idea to use RF alongside the commonly used algorithms in the challenging scenarios, such as long gaps, to figure out whether this superiority can be generalised.

2) It appears that, even after three levels of quality control process done by the PyFluxPro platform, the data are still noisy. This noisy data are an essential source of both uncertainty and inaccuracy of the outcome, regardless of the algorithm used to gap-fill the data. As a result, another level of quality control methods, such as Wavelets or Matrix Factorialisation, in addition to the current classical ones used by the PyFluxPro and other similar platforms, can probably improve the data quality and thereby improve the final imputation results.

3) For future researches, using recurrent neural networks (RNNs) instead of feedforward neural networks (FFNN) could improve the predictions. That is likely because RNNs help the model to consider temporal dynamic behaviour of time series, as unlike FFNN, wherein the activations flow only from the input layer to the output layer, RNNs also have neuron connections pointing backwards (Géron, 2019). This demand to an algorithm capable of considering time has been mentioned in previous researches as one of the reasons why testing the new algorithms is needed (Richardson and Hollinger, 2007).

3) Developing ensemble models using algorithms with different weaknesses and strengths may also enhance the results where a single algorithm shows performance deficiency.

6. Data availability

The data were used in this research are available through the following sources: The L3 and L4 data are accessible from the OzFlux data portal (http://data.ozflux.org.au/portal). Current ACCESS-R and data are available from the BoM OPeNDAP server (https://www.opendap.org/). Likewise, the data coming from the BoM AWS are accessible from (http://www.bom.gov.au/climate/data). Lastly, the BIOS2 data are accessible from the ECMWF datasets portal (https://www.ecmwf.int/en/forecasts/datasets). All data used in this research are available in this repository address: (https://research-repository.uwa.edu.au/en/datasets/a-comparison-of-gap-filling-algorithms-for-eddy-covariance-fluxes); DOI: 10.26182/5f292ee80a0c0.

Deleted: .

Moved up [3]: The same solution as number 4 is suggested for soil moisture estimation, as the behaviour of the system on sunny days is utterly different from its conduct during the rainy periods. Moreover, the system memory and the antecedent condition are undeniable features associated with soil moisture (Ogle et al., 2015). Therefore, using the models that are capable of addressing these considerations are more likely to improve the estimations.

Deleted: 4) Given that some of the environmental drivers affect the Fc differently during the day versus night, separating the diurnal and nocturnal datasets to train the algorithms possibly entails an improvement in the outcome. Mainly because of the u* threshold filtering and other problems associated with the nocturnal period, the portion of diurnal data is generally, by far, outweighs the nocturnal data portion, which potentially leads to a bias in the algorithm. ¶
5)

Moved up [2]: Given that some of the environmental drivers affect the Fc differently during the day versus night, separating the diurnal and nocturnal datasets to train the algorithms possibly entails an improvement in the outcome. Mainly because of the u* threshold filtering and other problems associated with the nocturnal period, the portion of diurnal data is generally, by far, outweighs the nocturnal data portion, which potentially leads to a bias in the algorithm.

Deleted: ¶

Author contributions. The ideas for this study originated in discussions with A. Mahabbati, J. Beringer, and M. Leopold. A. Mahabbati carried out the analysis, supported by I. McHugh and P. Isaac. The paper was prepared with contributions from all authors.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The authors would like to acknowledge Terrestrial Ecosystems Research Network (TERN) (www.tern.gov.au) and the OzFlux Network as a part of TERN for supporting the grants and providing the required data, respectively. A. Mahabbati also personally thanks Prajwal Kalfe, Caroline Johnson and Cacilia Ewenz for their support as regards Python programming, English academic writing and PyFluxPro technical issues.

References

Allison, P. D.: Multiple Imputation for Missing Data: A Cautionary Tale, Sociol. Methods Res., 28(3), 301–309, doi:10.1177/0049124100028003003, 2000.

Altman, D. G. and Bland, J. M.: Missing data, Br. Med. J., 334(7590), 424, doi:10.1136/bmj.38977.682025.2C, 2007.

Aubinet, M., Grelle, A., Ibrom, A., Rannik, Ü., Moncrieff, J., Foken, T., Kowalski, A. S., Martin, P. H., Berbigier, P., Bernhofer, C., Clement, R., Elbers, J., Granier, A., Grünwald, T., Morgenstern, K., Pilegaard, K., Rebmann, C., Snijders, W., Valentini, R. and Vesala, T.: Estimates of the Annual Net Carbon and Water Exchange of Forests: The EUROFLUX Methodology, Adv. Ecol. Res., 30, 113–175, doi:10.1016/S0065-2504(08)60018-5, 1999.

Aubinet, M., Vesala, T. and Papale, D.: Eddy Covariance: A Practical Guide to Measurement and Data Analysis., 2012a.

Aubinet, M., Vesala, T. and Papale, D.: Eddy Covariance., 2012b.

Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, C., Davis, K., Evans, R., Fuentes, J., Goldstein, A., Katul, G., Law, B., Lee, X., Malhi, Y., Meyers, T., Munger, W., Oechel, W., Paw, U. K. T., Pilegaard, K., Schmid, H. P., Valentini, R., Verma, S., Vesala, T., Wilson, K. and Wofsy, S.: FLUXNET: A New Tool to Study the Temporal and Spatial Variability of Ecosystem-Scale Carbon Dioxide, Water Vapor, and Energy Flux Densities, Bull. Am. Meteorol. Soc., 82(11), 2415–2434, doi:10.1175/1520-0477(2001)082<2415:FANTTS>2.3.CO;2, 2001.

Baltagi, B.: Econometric analysis of panel data, [online] Available from: http://www.sidalc.net/cgi-bin/wxis.exe/?IsisScript=book2.xis&method=post&formato=2&cantidad=1&expresion=mfn=001143 (Accessed 13 March 2018), 1995.

Barr, A. G., Black, T. A., Hogg, E. H., Kljun, N., Morgenstern, K. and Nesic, Z.: Inter-annual variability in the leaf area index of a boreal aspen-hazelnut forest in relation to net ecosystem production, Agric. For. Meteorol., 126(3–4), 237–255, doi:10.1016/J.AGRFORMET.2004.06.011, 2004.

Barr, A. G., Richardson, A. D., Hollinger, D. Y., Papale, D., Arain, M. A., Black, T. A., Bohrer, G., Dragoni, D., Fischer, M. L., Gu, L., Law, B. E., Margolis, H. A., Mccaughey, J. H., Munger, J. W., Oechel, W. and Schaeffer, K.: Use of change-point detection for friction-velocity threshold evaluation in eddy-covariance studies, Agric. For. Meteorol., 171–172, 31–45, doi:10.1016/j.agrformet.2012.11.023, 2013.

Bennett, N. D., Croke, B. F. W., Guariso, G., Guillaume, J. H. A., Hamilton, S. H., Jakeman, A. J., Marsili-Libelli, S., Newham, L. T. H., Norton, J. P., Perrin, C., Pierce, S. A., Robson, B., Seppelt, R., Voinov, A. A., Fath, B. D. and Andreassian, V.: Characterising performance of environmental models, Environ. Model. Softw., 40, 1–20, doi:10.1016/j.envsoft.2012.09.011, 2013.

Beringer, J., Hutley, L. B., McHugh, I., Arndt, S. K., Campbell, D., Cleugh, H. A., Cleverly, J., De Dios, V. R., Eamus, D., Evans, B., Ewenz, C., Grace, P., Griebel, A., Haverd, V., Hinko-Najera, N., Huete, A., Isaac, P., Kanniah, K., Leuning, R., Liddell, M. J., MacFarlane, C., Meyer, W., Moore, C., Pendall, E., Phillips, A., Phillips, R. L., Prober, S. M., Restrepo-Coupe, N., Rutledge, S., Schroder, I., Silberstein, R., Southall, P., Sun Yee, M., Tapper, N. J., Van Gorsel, E., Vote, C., Walker, J. and Wardlaw, T.: An introduction to the Australian and New Zealand flux tower network - OzFlux, Biogeosciences, 13(21), 5895–5916, doi:10.5194/bg-13-

5895-2016, 2016a.

Beringer, J., McHugh, I. and KLJUN, N.: Dynamic INtegrated Gap filling and partitioning for Ozflux (DINGO), Biogeosciences Discuss., OzFlux spe(In prep), 1457–1460, doi:doi:10.5194/bg-2016-188, 2016b.

Beringer, J., McHugh, I., Hutley, L. B., Isaac, P. and Kljun, N.: Technical note: Dynamic INtegrated Gap-filling and partitioning for OzFlux (DINGO), Biogeosciences, 14(6), 1457–1460, doi:10.5194/bg-14-1457-2017, 2017.

Burba, G. and Anderson, D.: A brief practical guide to eddy covariance flux measurements: principles and workflow examples for scientific and industrial applications. [online] Available from: https://books.google.com/books?hl=en&lr=&id=mCsI1_8GdrIC&oi=fnd&pg=PA6&dq=A+Brief+Practical+Guide+to+Eddy+Covariance+Flux+Measurements&ots=TKTg25Yq5X&sig=eBYc819N7Jh3gNhJInfEL1e40eM (Accessed 11 February 2020), 2010.

Chen, T. and Guestrin, C.: XGBoost: A scalable tree boosting system, Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 13-17-Augu, 785–794, doi:10.1145/2939672.2939785, 2016.

Cleverly, J., Boulain, N., Villalobos-Vega, R., Grant, N., Faux, R., Wood, C., Cook, P. G., Yu, Q., Leigh, A. and Eamus, D.: Dynamics of component carbon fluxes in a semi-arid *Acacia* woodland, central Australia, J. Geophys. Res. Biogeosciences, 118(3), 1168–1185, doi:10.1002/jgrg.20101, 2013.

Devore, J. L.: Probability and Statistics for Engineering and the Sciences., Biometrics, 47(4), 1638, doi:10.2307/2532427, 1991.

Dragoni, D., Schmid, H. P., Grimmond, C. S. B. and Loescher, H. W.: Uncertainty of annual net ecosystem productivity estimated using eddy covariance flux measurements, J. Geophys. Res., 112(D17), D17102, doi:10.1029/2006JD008149, 2007.

Dreyfus, S. E.: Artificial neural networks, back propagation, and the kelley-bryson gradient procedure, J. Guid. Control. Dyn., 13(5), 926–928, doi:10.2514/3.25422, 1990.

Drucker, H., Surges, C. J. C., Kaufman, L., Smola, A. and Vapnik, V.: Support vector regression machines, in Advances in Neural Information Processing Systems, vol. 1, pp. 155–161., 1997.

Falge, E., Baldocchi, D., Olson, R., Anthoni, P., Aubinet, M., Bernhofer, C., Burba, G., Ceulemans, R., Clement, R., Dolman, H., Granier, A., Gross, P., Grünwald, T., Hollinger, D., Jensen, N. O., Katul, G., Keronen, P., Kowalski, A., Lai, C. T., Law, B. E., Meyers, T., Moncrieff, J., Moors, E., Munger, J. W., Pilegaard, K., Rannik, Ü., Rebmann, C., Suyker, A., Tenhunen, J., Tu, K., Verma, S., Vesala, T., Wilson, K. and Wofsy, S.: Gap filling strategies for defensible annual sums of net ecosystem exchange, Agric. For. Meteorol., 107(1), 43–69, doi:10.1016/S0168-1923(00)00225-2, 2001.

Farley, B. G. and Clark, W. A.: Simulation of self-organizing systems by digital computer, IRE Prof. Gr. Inf. Theory, 4(4), 76–84, doi:10.1109/TIT.1954.1057468, 1954.

Freedman, D. A.: Statistical Models: Theory and Practice. Cambridge University Press - 2nd edition. [online] Available from: https://www.cambridge.org/au/academic/subjects/statistics-probability/statistical-theory-and-methods/statistical-models-theory-and-practice-2nd-edition?format=PB (Accessed 21 March 2020), 2009.

Friedman, J. H.: Stochastic gradient boosting, Comput. Stat. Data Anal., 38(4), 367–378, doi:10.1016/S0167-9473(01)00065-2, 2002.

Gani, A., Mohammadi, K., Shamshirband, S., Altameem, T. A., Petković, D. and Ch, S.: A combined method to estimate wind speed distribution based on integrating the support vector machine with firefly algorithm, Environ. Prog. Sustain. Energy, 35(3), 867–875, doi:10.1002/ep.12262, 2016.

Géron, A.: Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. [online] Available from: https://books.google.com.au/books?hl=en&lr=&id=HHetDwAAQBAJ&oi=fnd&pg=PP1&dq=hands-on+machine+learning+with+&ots=0KvfZqlgOo&sig=5tH2IHRsUaTMTy6CfQ6lw3UDKa4 (Accessed 7 February 2020), 2019.

Hagen, S. C., Braswell, B. H., Linder, E., Frolking, S., Richardson, A. D. and Hollinger, D. Y.: Statistical uncertainty of eddy flux - Based estimates of gross ecosystem carbon exchange at Howland Forest, Maine, J. Geophys. Res. Atmos., 111(8), 1–12, doi:10.1029/2005JD006154, 2006.

Harrell, F. E.: Regression Modeling Strategies: With Applications to Linear Models, Logistic, in books.google.nl. [online] Available from: https://books.google.com.au/books?hl=en&lr=&id=94RgCgAAQBAJ&oi=fnd&pg=PR7&dq=regression+modeling+strategies+frank+harrell&ots=ZAt4RsaS1r&sig=mikE1s9G4IXzqZKEie-iVA9GTV0&redir_esc=y#v=onepage&q=regression modeling strategies frank harrell&f=false (Accessed 11 February 2020), 2014.

Harvey, A. C. and Peters, S.: Estimation procedures for structural time series models, J. Forecast., 9(2), 89–108, doi:10.1002/for.3980090203, 1990.

Haverd, V., Briggs, P., Trudinger, C., Nieradzik, L. and Canadell, P.: BIOS2 – Frontier Modelling of the Australian Carbon and Water Cycles, 2015.

Ho, T. K.: Random decision forests, Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, 1, 278–282, doi:10.1109/ICDAR.1995.598994, 1995.

Ho, T. K.: 00709601.Pdf, , 20(8), 832–844, 1998.

Hollinger, D. Y., Goltz, S. M., Davidson, E. A., Lee, J. T., Tu, K. and Valentine, H. T.: Seasonal patterns and environmental control of carbon dioxide and water vapour exchange in an ecotonal boreal forest, Glob. Chang. Biol., 5(8), 891–902, doi:10.1046/j.1365-2486.1999.00281.x, 1999.

Hsiao, C., Hashem Pesaran, M. and Kamil Tahmiscioglu, A.: Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods, J. Econom., 109(1), 107–150, doi:10.1016/S0304-4076(01)00143-9, 2002.

Hui, D., Wan, S., Su, B., Katul, G., Monson, R. and Luo, Y.: Gap-filling missing data in eddy covariance measurements using multiple imputation (MI) for annual estimations, Agric. For. Meteorol., 121(1–2), 93–111, doi:10.1016/S0168-1923(03)00158-8, 2004.

Hutley, L. B., Leuning, R., Beringer, J. and Cleugh, H. a: The utility of the eddy covariance technique as a tool in carbon accounting: tropical savanna as a case study, Aust. J. Bot., 53, 663–675, 2005.

Isaac, P., Cleverly, J., McHugh, I., Van Gorsel, E., Ewenz, C. and Beringer, J.: OzFlux data: Network integration from collection to curation, Biogeosciences, 14(12), 2903–2928, doi:10.5194/bg-14-2903-2017, 2017.

Izady, A., Davary, K., Alizadeh, A., Moghaddam Nia, A., Ziaei, A. N. and Hasheminia, S. M.: Application of NN-ARX Model to Predict Groundwater Levels in the Neishaboor Plain, Iran, Water Resour. Manag., 27(14), 4773–4794, doi:10.1007/s11269-013-0432-y, 2013.

Izady, A., Abdalla, O. and Mahabbati, A.: Dynamic panel-data-based groundwater level prediction and decomposition in an arid hardrock–alluvium aquifer, Environ. Earth Sci., 75(18), 1–13, doi:10.1007/s12665-016-6059-6, 2016.

Jerome H. Friedman: Greedy Function Approximation: A Gradient Boosting Machine on JSTOR, Ann. Stat., 29, 1189–1232 [online] Available from: https://www.jstor.org/stable/2699986?seq=1#metadata_info_tab_contents (Accessed 27 August 2019), 2001.

Kang, H.: The prevention and handling of the missing data, Korean J. Anesthesiol., 64(5), 402–406, doi:10.4097/kjae.2013.64.5.402, 2013.

Kim, Y., Johnson, M. S., Knox, S. H., Black, T. A., Dalmagro, H. J., Kang, M., Kim, J. and 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., 26(3), 1499–1518, doi:10.1111/gcb.14845, 2020.

Kock, N. and Gaskins, L.: Simpson's paradox, moderation and the emergence of quadratic relationships in path models: an information systems illustration, Int. J. Appl. Nonlinear Sci., 2(3), 200, doi:10.1504/ijans.2016.077025, 2016.

Kunwor, S., Starr, G., Loescher, H. W. and Staudhammer, C. L.: Preserving the variance in imputed eddy-covariance measurements: Alternative methods for defensible gap filling, Agric. For. Meteorol., 232, 635–649, doi:10.1016/j.agrformet.2016.10.018, 2017.

Law, B. E., Falge, E., Gu, L., Baldocchi, D. D., Bakwin, P., Berbigier, P., Davis, K., Dolman, A. J., Falk, M., Fuentes, J. D., Goldstein, A., Granier, A., Grelle, A., Hollinger, D., Janssens, I. A., Jarvis, P., Jensen, N. O., Katul, G., Mahli, Y., Matteucci, G., Meyers, T., Monson, R., Munger, W., Oechel, W., Olson, R., Pilegaard, K., Paw U H, K. T., Thorgeirsson, H., Valentini, R., Verma, S., Vesala, T., Wilson, K. and Wofsy, S.: Jourassess2, Agric. For. Meteorol., 113(113), 97–120, 2002.

Lee, X., Fuentes, J. D., Staebler, R. M. and Neumann, H. H.: Long-term observation of the atmospheric exchange of CO2 with a temperate deciduous forest in southern Ontario, Canada, J. Geophys. Res. Atmos., 104(D13), 15975–15984, doi:10.1029/1999JD900227, 1999.

Little, R. J. A.: Statistical analysis with missing data, 2nd ed., edited by D. B. Rubin, Wiley, Hoboken, N.J., 2002.

Mahabbati, A., Izady, A., Mousavi Baygi, M., Davary, K. and Hasheminia, S. M.: Daily soil temperature modeling using 'panel-data' concept, J. Appl. Stat., 44(8), 1385–1401, doi:10.1080/02664763.2016.1214240, 2017.

Menzer, O., Moffat, A. M., Meiring, W., Lasslop, G., Schukat-Talamazzini, E. G. and Reichstein, M.: Random errors in carbon and water vapor fluxes assessed with Gaussian Processes, Agric. For. Meteorol., 178–179, 161–172, doi:10.1016/j.agrformet.2013.04.024, 2013.

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., Falge, E., Gove, J. H., Heimann, M., Hui, D., Jarvis, A. J., Kattge, J., Noormets, A. and Stauch, V. J.: Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes, Agric. For. Meteorol., 147(3–4), 209–232, doi:10.1016/j.agrformet.2007.08.011, 2007.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., Verbeke, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A. and Verbeke, G.: Handbook of Missing Data Methodology, Chapman and Hall/CRC., 2014.

Ogle, K., Barber, J. J., Barron-Gafford, G. A., Bentley, L. P., Young, J. M., Huxman, T. E., Loik, M. E. and Tissue, D. T.: Quantifying ecological memory in plant and ecosystem processes, Ecol. Lett., 18(3), 221–235, doi:10.1111/ele.12399, 2015.

Papale, D. and Valentini, R.: A new assessment of European forests carbon exchanges by eddy fluxes and artificial neural network spatialization, Glob. Chang. Biol., 9(4), 525–535, doi:10.1046/j.1365-2486.2003.00609.x, 2003.

Pilegaard, K., Hummelshøj, P., Jensen, N. O. and Chen, Z.: Two years of continuous CO2 eddy-flux measurements over a Danish beech forest, Agric. For. Meteorol., 107(1), 29–41, doi:10.1016/S0168-1923(00)00227-6, 2001.

Reichle, R. H., Koster, R. D., Dong, J. and Berg, A. A.: Global soil moisture from satellite observations, land surface models, and ground data: Implications for data assimilation, J. Hydrometeorol., 5(3), 430–442, doi:10.1175/1525-7541(2004)005<0430:GSMFSO>2.0.CO;2, 2004.

Reichstein, M., Falge, E., Baldocchi, D., Papale, D., Aubinet, M., Berbigier, P., Bernhofer, C., Buchmann, N., Gilmanov, T., Granier, A., Grünwald, T., Havránková, K., Ilvesniemi, H., Janous, D., Knohl, A., Laurila, T., Lohila, A., Loustau, D., Matteucci, G., Meyers, T., Miglietta, F., Ourcival, J. M., Pumpanen, J., Rambal, S., Rotenberg, E., Sanz, M., Tenhunen, J., Seufert, G., Vaccari, F., Vesala, T., Yakir, D. and Valentini, R.: On the separation of net ecosystem exchange into assimilation and ecosystem respiration: Review and improved algorithm, Glob. Chang. Biol., 11(9), 1424–1439, doi:10.1111/j.1365-2486.2005.001002.x, 2005.

Richardson, A. D. and Hollinger, D. Y.: A method to estimate the additional uncertainty in gap-filled NEE resulting from long gaps in the CO2 flux record, Agric. For. Meteorol., 147(3–4), 199–208, doi:10.1016/j.agrformet.2007.06.004, 2007.

Richardson, A. D., Braswell, B. H., Hollinger, D. Y., Burman, P., Davidson, E. A., Evans, R. S., Flanagan, L. B., Munger, J. W., Savage, K., Urbanski, S. P. and Wofsy, S. C.: Comparing simple respiration models for eddy flux and dynamic chamber data, Agric. For. Meteorol., 141(2–4), 219–234, doi:10.1016/J.AGRFORMET.2006.10.010, 2006.

Richardson, A. D., Aubinet, M., Barr, A. G., Hollinger, D. Y., Ibrom, A., Lasslop, G. and Reichstein, M.: Uncertainty Quantification, in Eddy Covariance, pp. 173–209., 2012.

Sahoo, A. K., Dirmeyer, P. A., Houser, P. R. and Kafatos, M.: A study of land surface processes using land surface models over the Little River Experimental Watershed, Georgia, J. Geophys. Res. Atmos., 113(20), doi:10.1029/2007JD009671, 2008.

Scanlon, T. M. and Kustas, W. P.: Partitioning carbon dioxide and water vapor fluxes using correlation analysis, Agric. For. Meteorol., 150(1), 89–99, doi:10.1016/j.agrformet.2009.09.005, 2010.

Scanlon, T. M. and Sahu, P.: On the correlation structure of water vapor and carbon dioxide in the atmospheric surface layer: A basis for flux partitioning, Water Resour. Res., 44(10), doi:10.1029/2008WR006932, 2008.

Staebler, M.: Long-term observation of the atmospheric exchange of CO2 with a temperate deciduous forest in southern Ontario , Canada ecosystem (net ecosystem production turbulence is turbulent, Data Process., 104, 975–984, 1999.

Tannenbaum, C. E.: The empirical nature and statistical treatment of missing data., Diss. Abstr. Int. Sect. A Humanit. Soc. Sci., 70(10-A), 3825 [online] Available from: http://gateway.proquest.com/openurl?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&res_dat=xri:pqdiss&rft_dat=xri:pqdiss:3381876%5Cnhttp://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc7&NEWS=N&AN=2010-99071-044, 2010.

Taylor, S. J. and Letham, B.: Business Time Series Forecasting at Scale, , doi:10.7287/peerj.preprints.3190v2, 2017.

Taylor, S. J. and Letham, B.: Forecasting at Scale, Am. Stat., 72(1), 37–45, doi:10.1080/00031305.2017.1380080, 2018.

Tenhunen, J. D., Valentini, R., Köstner, B., Zimmermann, R. and Granier, A.: Variation in forest gas exchange at landscape to continental scales, Ann. des Sci. For., 55(1–2), 1–11, doi:10.1051/forest:19980101, 1998.

Wooldridge, J. M.: Econometric Analysis of Cross Section and Panel Data., 2008.

Ye, J., Chow, J.-H., Chen, J. and Zheng, Z.: Stochastic gradient boosted distributed decision trees, in Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09, p. 2061, ACM Press, New York, New York, USA., 2009.

Zhao, X. and Huang, Y.: A comparison of three gap filling techniques for eddy covariance net carbon fluxes in short vegetation ecosystems, Adv. Meteorol., 2015, 1–12, doi:10.1155/2015/260580, 2015.

Zou, H. and Hastie, T.: Regularization and variable selection via the elastic net. [online] Available from: https://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=22250F01CC77D55C54B6BAFF4512C9E3?doi=10.1.1.124.4696&rep=rep1&type=pdf (Accessed 28 August 2019), 2005

Formatted: Indent: First line: 0 cm

Deleted: ¶