https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

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 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

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020


Geoscientific Instrumentation Methods and Data Systems Discussions

37 those of machine learning (ML), except the random forest algorithm estimating the ground heat flux
38 (RMSEs of 30.17 and 34.93 for RF and CLR respectively). For the major fluxes, though, ML algorithms
39 showed superiority (9 % less RMSE on average), except the Support Vector Regression (SVR), which
40 provided significant bias in its estimations. Even though ANNs, random forest (RF) and extreme
41 gradient boost (XGB) showed close performance in gap-filling of the major fluxes, RF provided more
42 consistent results with less bias, relatively. The results indicated that there is no single algorithm
43 which outperforms in all situations and therefore, but RF is a potential alternative for the ANNs as
44 regards flux gap-filling.

45

46 1. Introduction

47 To address the global challenges of climatological and ecological changes, environmental
48 scientists and policymakers are demanding data that are continuous in time and space. Besides, there
49 is a need for quantifying and reducing uncertainties in such data, including observations of carbon,
50 water and energy exchanges that are crucial components in national/international flux networks and
51 global earth observing systems. Satellites partially fill this gap as they provide excellent spatial
52 coverage but at a limited temporal resolution, and not measured at the point. As such, high-quality
53 long-term site observations of ecosystem process and fluxes are needed that are continuous in time
54 and space. The global eddy covariance (EC) flux tower networks (FLUXNET), consisted of its regional
55 counterparts (i.e. AmeriFlux, EUROFLUX, OzFlux, etc.), was established in the late 1990s to address
56 the global demand for such information (Beringer et al., 2016a; Hollinger et al., 1999; Tenhunen et al.,
57 1998). Despite the capability of EC to frequently validate process modelling analyses, field surveys
58 and remote sensing assessments (Hagen et al., 2006), there are some serious concerns regarding the
59 challenges associated with the technique, e.g. data gaps and uncertainties. Hence, filling data gaps
60 and reducing uncertainties through better gap-filling techniques are highly needed.

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

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020


Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

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).

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., 2012). 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., 2012). 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., 2012; 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 a couple of 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}. This result is similar to the findings of Richardson & Hollinger (2007) who found 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.

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

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020


Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

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., 2012). 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., 2012). 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., 2012; 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 CO2 flux in the FLUXNET (Aubinet et al., 2012). 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

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020


Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

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 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) evaluate the performance of traditional and new gap-filling techniques for the OzFlux Network, 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 eight different algorithms (Extreme Gradient Boost (XGB), Random Forest Algorithm (RF), Artificial Neural Networks (ANNs), 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 environmental drivers and the major fluxes. 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.

2. Materials and methods

To address the first objective of this research, data of nine different window lengths were considered to train and test the algorithms, i.e. 1, 5, 10, 20, 30, 60, 90, 180 and 365 days. To address the second objective, we chose eight 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. 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 form 2011 to 2013, with a time resolution of 30 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.

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

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

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 2011 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, respectively.

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

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

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)
Fluxes:	
Fc	CO_2 flux (μmol m^{-2} s^{-1})
Fh	Sensible heat flux (W m^{-2})
Fe	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, Fh and Fe, eight drivers were used as follows: Ta, Ws, Sws, Fg, VPD, Fn, q and Ts based on trial and error. 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 gird search, and the number of nodes, layers, irritations, etc. were chosen therefor.

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

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020

Geoscientific Instrumentation Methods and Data Systems Discussions

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 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 a standard method like ANNs, to the newer algorithms which rarely or never been used in the field, such as Extreme Gradient Boosting and panel data. 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.

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)

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.

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020


Geoscientific Instrumentation Methods and Data Systems Discussions

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:

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

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

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):

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

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:

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020


Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

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

296

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

301

302 **Panel data (PD)**

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

$$y_{it} = \alpha + \beta X_{it} + u_{it} \qquad i = 1, 2, ..., N; \;\; t = 1, 2, ..., T \quad (4)$$

$$y_{it} = \alpha + \beta X_{it} + \mu_i + \lambda_t \quad (5)$$

308 where i and t denote the cross-section and time series dimension in a row, y is a dependent-variable
309 vector, X is an independent variable matrix, α is a scalar, β is the coefficient of the independent-
310 variable matrix, μ_i is the unobservable individual-specific effect, and λ_t is the unobservable time-
311 specific effect. Panel data abilities to provide a holistic analysis of different individuals, as well as
312 determining the specific impact of every single time caused its superiority over CLR.

313 **Extreme Gradient Boost (XGB)**

314 Extreme gradient boost is a reinforced method of Gradient Boost introduced in 1999 that works based
315 on parallel boosted decision trees and similar to RF can be used for a variety of data processing
316 purposes including classification and regression (Friedman, 2002; Jerome H. Friedman, 2001; Ye et al.,
317 2009). XGB method is resistive to over-fitting and provides a robust, portable and scalable algorithm
318 for large-scale boosting decision-trees-based techniques.
319 sklearn.ensemble.GradientBoostingRegressor was used to apply this method in Python, and its
320 hyperparameters were chosen based on grid search as follows: {'learning_rate': 0.001, 'max_depth': 8,
321 'reg_alpha': 0.1, 'subsample': 0.5}.

322

323 **The Prophet Forecasting Model (FBP)**

324 The Prophet Forecasting Model, also known as "prophet", is a time series forecasting model
325 developed by Facebook to manage the common features of business time series and designed to have

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020


Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

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):

$$y(t) = g(t) + s(t) + h(t) \quad (6)$$

where g(t) is the trend function, which models mom-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.

2.3. The gap scenarios

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.

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):

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020


Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

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

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

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

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

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

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

where o_i and p_i are individual measured and predicted values respectively, $\bar{o}$ and $\overline{p}$ 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.

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. The algorithms, however, showed different levels of sensitivity to training/testing window length, e.g. the ANNs showed less sensitivity, whereas the FBP showed

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

the most sensitivity (Figure 1). 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.

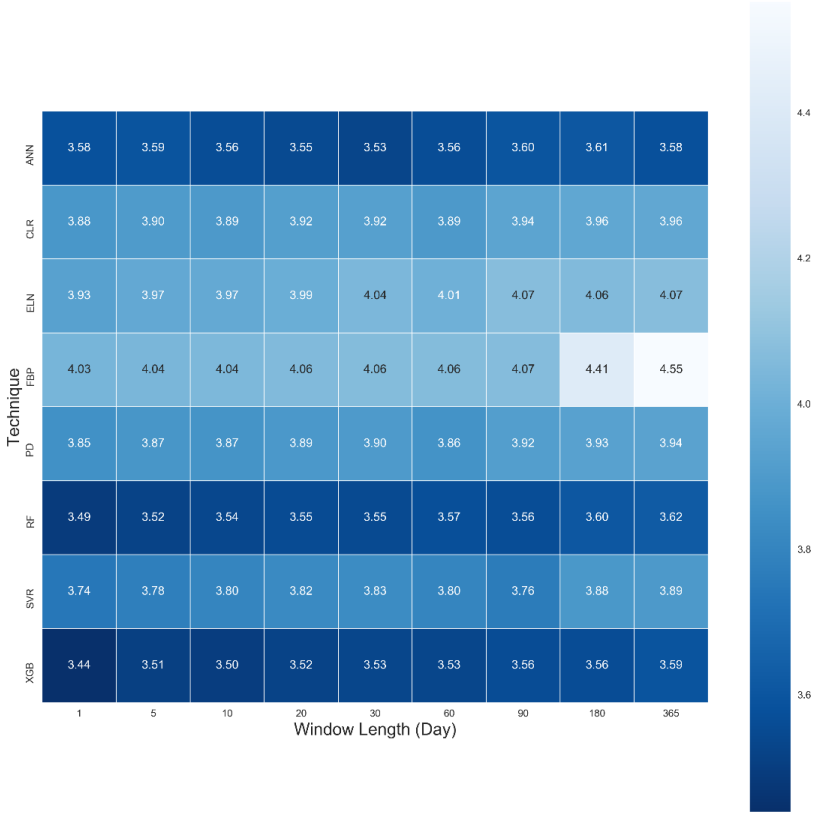
Table 4. The average amounts of performance metrics for each gap-filling algorithm regarding Fc, which includes all window lengths and sites, ranked by RMSE using the Tukey's HSD test at the level of 5 per cent.

Algorithm	Mean RMSE	Mean R^2	Mean MBE	Mean IoAd	Mean VR
XGB	3.53 [a]	0.56	-0.44	0.89	0.59
RF	3.56 [a]	0.54	-0.38	0.90	0.70
ANNs	3.57 [a]	0.52	-0.34	0.89	0.68
SVR	3.81 [b]	0.47	-0.33	0.86	0.79
PD	3.89 [b]	0.45	-0.36	0.80	0.53
CLR	3.92 [b,c]	0.46	-0.37	0.80	0.54
ELN	4.01 [c]	0.40	-0.38	0.72	0.37
FBP	4.15 [d]	0.44	-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) that leads to better performance. 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 algorithms, and there is no significant difference amongst the appropriate algorithms. Nonetheless, it is worth 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 ANN, as the superiority of RF in gap-filling of methane flux has recently been claimed by (Kim et al., 2020).

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020

Geoscientific Instrumentation Methods and Data Systems Discussions



411

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

413

414 To address the first objectives of this paper, finding out the sensitivity of each algorithm to the
415 training and testing window length, we used the averaged RMSE, R^2 and MBE for each window length
416 and gap size, using the output of all algorithms for all sites (Table 5). The outcome illustrates that the
417 longer the window length got, the bigger the amounts of RMSE became. Yet, no such pattern was
418 recognisable for the R^2 and MBE, particularly for the window lengths equal to or shorter than 60 days.
419 As a result, based on our scenarios (using the same length for training window and gap size), choosing
420 any training windows longer than 60 days, i.e. 90, 180 and 365 days, made the performance of the

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020

Geoscientific Instrumentation Methods and Data Systems Discussions

algorithms worst. 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	3.74 [a]	0.49	-0.26
5-days	3.77 [a]	0.49	-0.31
10-days	3.77 [a]	0.49	-0.29
20-days	3.79 [a]	0.48	-0.31
30-days	3.80 [a]	0.48	-0.31
60-days	3.79 [a]	0.48	-0.35
90-days	3.81 [a]	0.48	-0.39
180-days	3.88 [a]	0.47	-0.41
365-days	3.90 [a]	0.46	-0.37

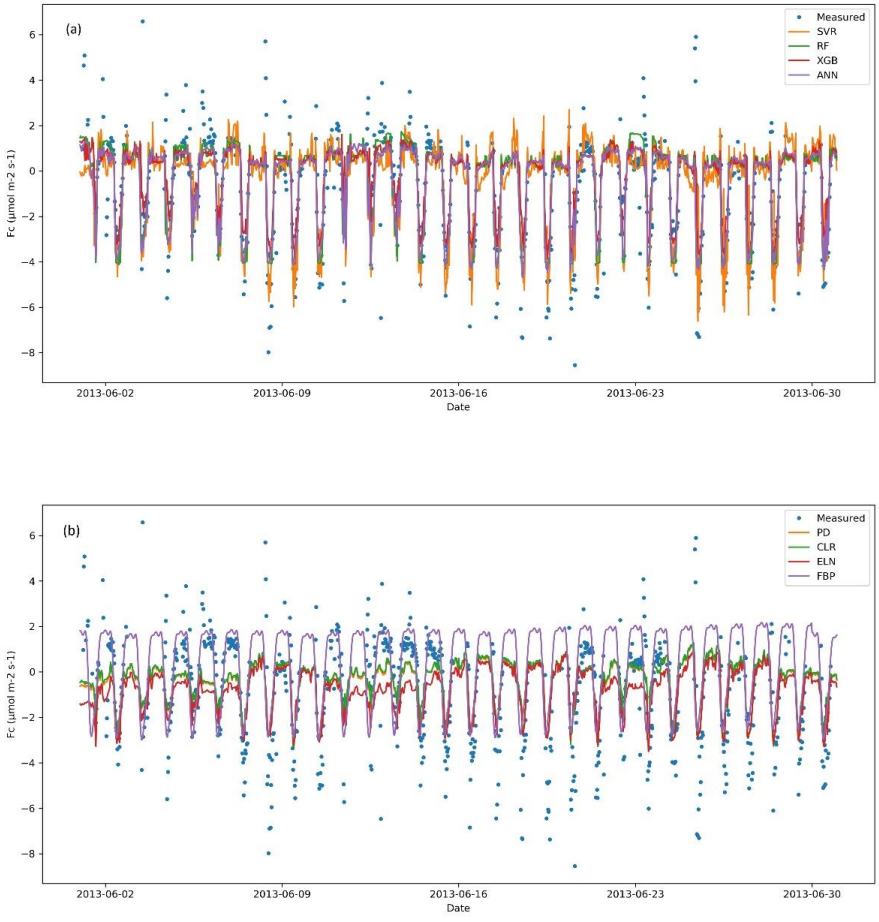
Besides, the metrics of the top three algorithms, XGB, RF and ANNs, did not show any sizeable difference for the window lengths shorter than 60 days. As such, finding the ideal window length, at least at this stage, is not distinctly noticeable and should rely on the local knowledge of each specific site. Nevertheless, as mentioned earlier, the ideal window generally cannot be longer than 60 days, unless for a monotonic ecosystem without a dramatic change during the year. According to the MBE values, 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 (results not shown). 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 -0.33 and -0.44.

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., 2012; 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 (Figure 2). 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 also recognisable in Figure 2. The variance ratios (VR) showed that SVR captured the extreme values of Fc better than the other algorithms, 0.79 on average. The XGB, on

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020

Geoscientific Instrumentation Methods and Data Systems Discussions

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



450

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

452 This substantial smaller VR calls into question the ability of XGB to provide a solid gap-filling for long
453 gaps. The linear algorithms, CLR, PD, and ELN, performed worse with respect to the VR compared
454 to the ML algorithms. The estimated versus measured values of Fc for Calperum during June 2013
455 (Figure 2) confirms the information achieved by the VR. Based on the figure, and VR of 0.79, the SVR
456 captured the extreme values of Fc the best, whereas the ELN, as expected, performed the worst (0.37).
457 Although the XGB (VR of 0.59) provided relatively well while estimating the maximum values
458 (respiration), it was not capable of assessing the minimum values, thereby provided a constant
459 overestimation of NEE during the day. The RF (VR of 0.70), in contrast, captured both negative and
460 positive extremes better than the XGB, while the performance of the ANNs (VR of 0.68) was

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020

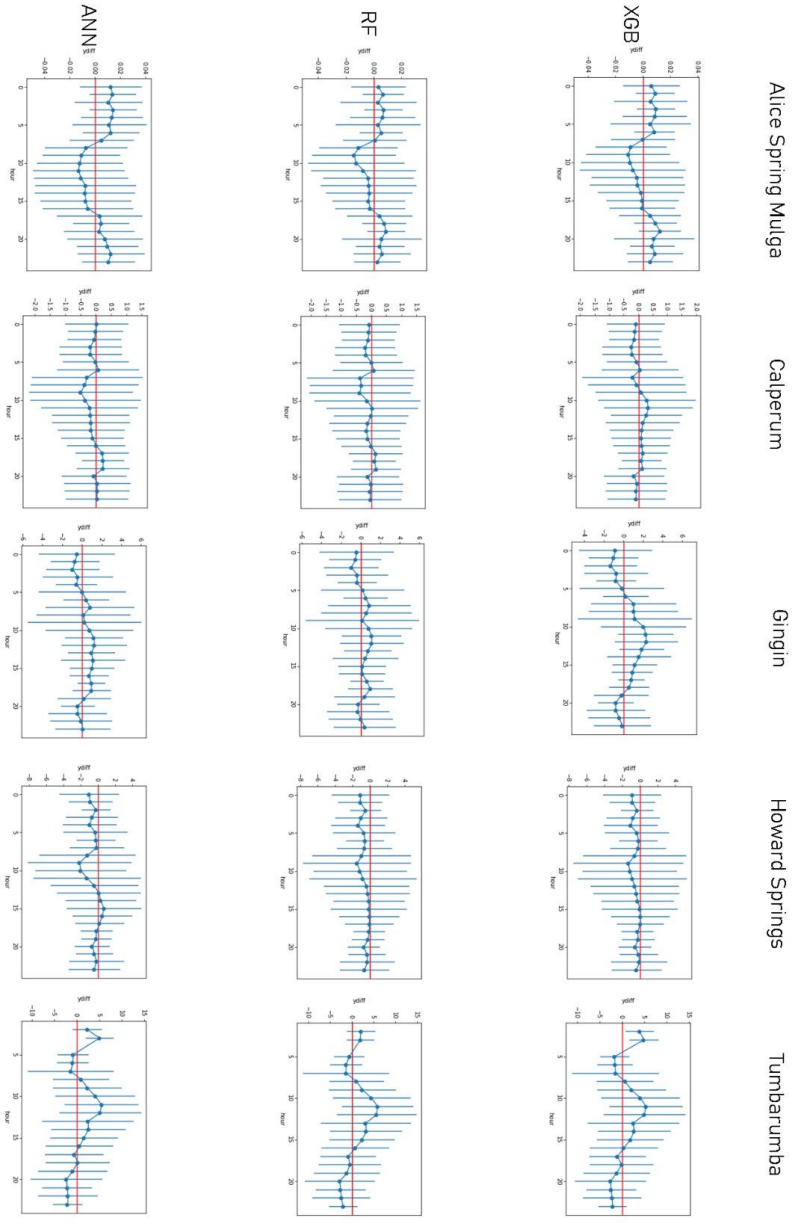

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

somewhere in between. The rest of the algorithms performed poorly, particularly during the night, except the FBP. It is noteworthy that CLR, PD, and ELN frequently predicted nocturnal photosynthesis.

Apart from the objectives of this paper, tracing the performance of gap-filling algorithms based on the hourly time step and seasonality has been as of the research interests. Thus, as an aside, the differences between the average of estimations and measured values, as well as the difference between the variances for the top three algorithms (XGB, RF and ANNs) were calculated for the 24-h and seasonal ranges. These algorithms showed different anomalies in different towers and hours of the day, except for Tumbarumba, where the patterns of anomalies were almost similar (Figure 3). The average variance of differences was slightly lower during the night while the largest values of anomalies usually occurred around noon, as expected due to the bigger variations of carbon uptake caused by photosynthesis. Here, the RF and XGB performed better than ANNs, with the curves closer to the basis, except for Tumbarumba. According to the seasonal anomalies, however, the algorithms showed more similarities and closer outcomes, particularly for Tumbarumba, where all three overestimated Fc values for the whole year (Figure 4). Similar to the 24-hour scale, the anomaly values varied from site to site based on the season and the algorithm. Although the performance of the algorithms was less here as against the daily scale, it seems that the XGB and RF still show superiority over the ANNs. Apart from Tumbarumba, XGB, RF and ANNs showed a significant bias during spring (July, August and September) in Howard Springs, when the site receives lower precipitation due to the dry season.

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU



491

492 *Figure 3. 24-h Anomalies of XGB, RF and ANNs based on the Fc average of differences, and associated variances between the*
493 *estimated and measured values for all towers during 2013.*

494

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020

Geoscientific Instrumentation Methods and Data Systems Discussions

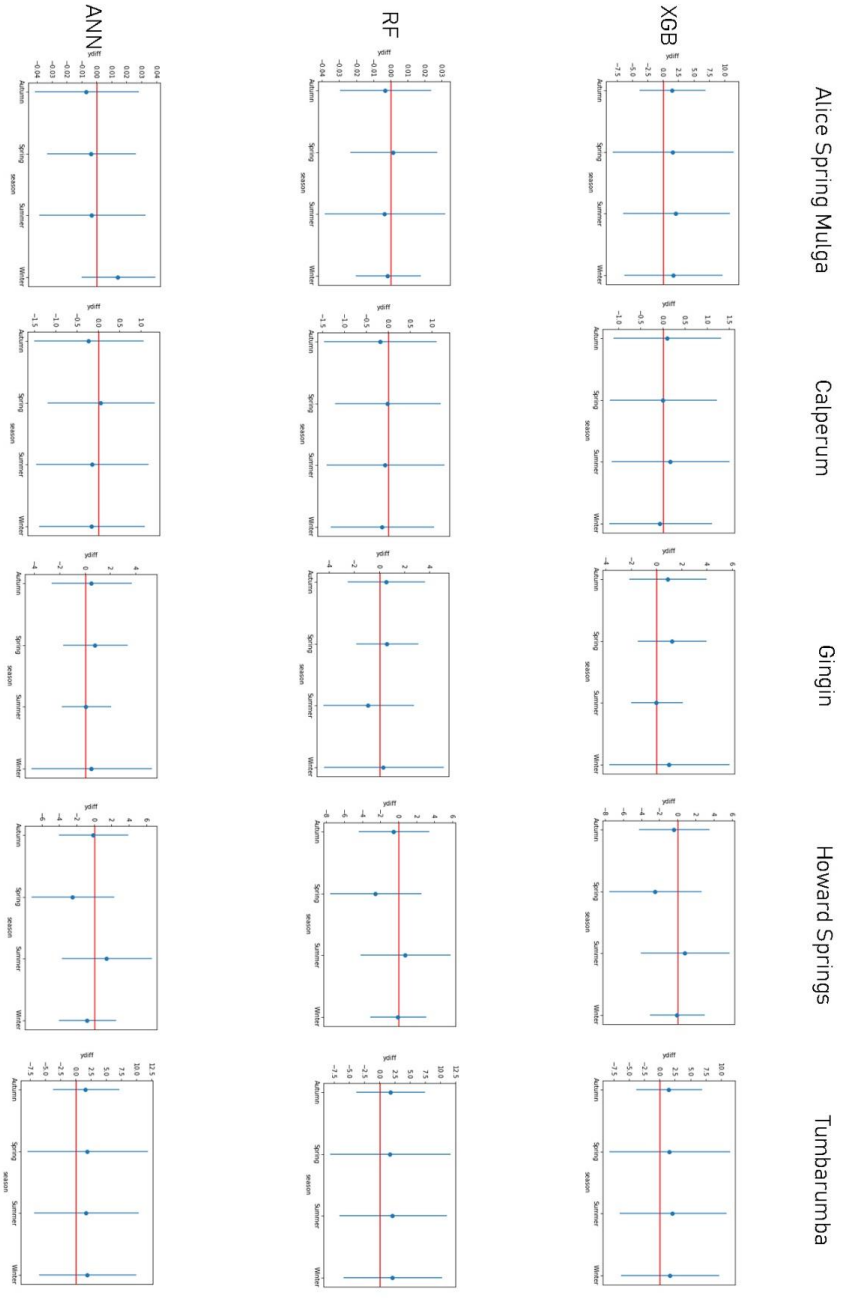


Figure 4. Seasonal anomalies of XGB, RF and ANNs based on the Fc averages of differences, and the associated variances between the estimated and measured values for all towers during 2013 (Jan, Feb and Mar as Summer, Apr, May and Jun as Autumn, Jul, Aug and Sep as winter).

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020

Geoscientific Instrumentation Methods and Data Systems Discussions

3.1.2 Fe

The performance of algorithms for Fe was similar to that for Fc regarding RMSE, MBE and R2, 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*). 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 window lengths of training and testing became shorter than 90 days, meaning that the performance of the algorithms did not vary considerably from a 60-day to a one-day window. 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 5), 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	37.27 [a]	0.69	-3.19
RF	37.98 [a]	0.68	-3.00
ANNs	40.62 [b]	0.61	-3.48
PD	42.45 [b,c]	0.58	-5.50
CLR	42.67 [b,c]	0.58	-5.95
Eln	43.48 [c]	0.53	-5.00
SVR	48.42 [d]	0.53	-21.08
FBP	49.46 [d]	0.44	2.03

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

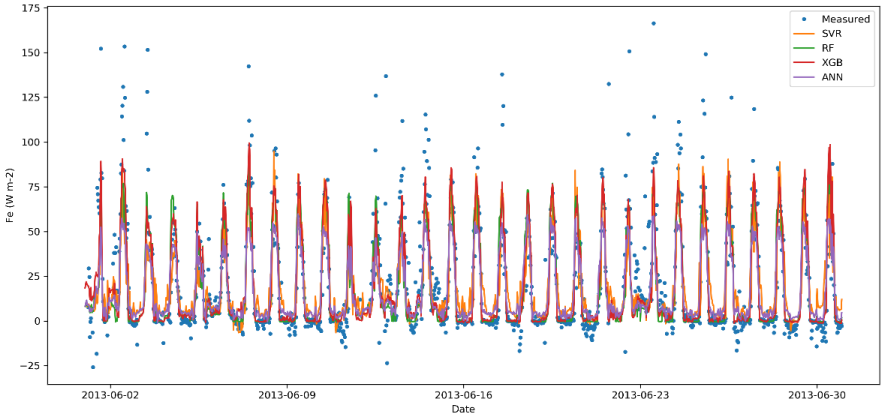


Figure 5. Measured vs estimated values of Fe for Calperum based on a 60-day window during June 2013

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, even though there was no significant difference amongst the XGB, ANNs and RF, the first one was considerably superior over the other five algorithms as regards the Tukey's HSD test. Like Fe, estimated values of Fh using SVR had a negative bias (Figure 6) because it was not able to provide appropriate estimations of Fh negative values. In contrast, the ANNs performed the best in capturing the minimum values, while the XGB and RF performed relatively well, close to each other. Despite this superiority in assessing the minimum amounts, ANNs 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	37.26 [a]	0.93	-1.00
RF	38.08 [a,b]	0.93	-1.35
ANNs	40.48 [a,b,c]	0.92	-0.41
PD	41.83 [b,c]	0.92	-0.27
CLR	42.14 [b,c]	0.92	-0.05
Eln	42.28 [b,c]	0.92	0.04
SVR	43.98 [c]	0.91	-8.28
FBP	67.19 [d]	0.74	1.25

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020

Geoscientific Instrumentation Methods and Data Systems Discussions

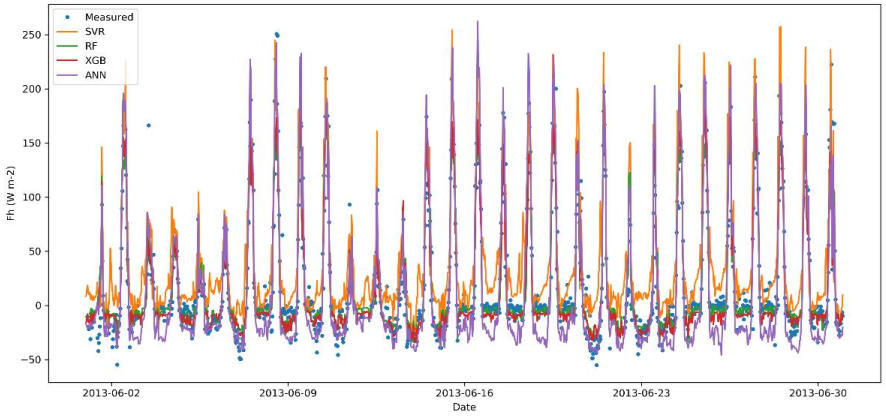


Figure 6. Measured vs estimated values of Fh for Calperum based on a 60-day window during June 2013

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 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 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 30.23 versus 35.24 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 changes dramatically during rainfall in a

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

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 0.43 and 0.25 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]	30.17
XGB [a, b]	30.70
ANNs [b, c]	30.86
SVR [c]	32.77
CLR [d]	34.93
PD [d]	34.94
ELN [d]	34.94
FBP [e]	39.10

4. Discussion

All algorithms 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 using training/gap windows of 60 days or shorter, while the worst results obtained for the most extended window, 365 days. Although most of the algorithms performed almost equally well in estimating of 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, which the RF provided more accurate and robust estimations. The ML algorithms, 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 SVR, FBP and all linear algorithms, i.e. the CLR, PD and ELN. 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 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).

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020


Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

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. 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. In estimating the meteorological drivers, the XGB did not show any superiority over the other algorithms, especially the linear ones. Hence, we do not recommend the XGB as an alternative to the current alternative algorithms, especially for long gaps. Nevertheless, because of its local superiorities, this algorithm is suitable to use in an ensemble model alongside the algorithms with different weakness points.

The RF was the best all-around algorithm amongst the eight algorithms used in this study, providing the best estimates of the fluxes (similar to XGB) but also captured both minimum and maximum values of Fc. Unlike the RF, all other algorithms generally struggled with estimating either minimum or maximum values of major fluxes, comparatively. It also provided the best results for Fg, where the linear algorithms did not perform well. 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.

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 90 days or longer. 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., 2012; Hagen et al., 2006; Kunwor et al., 2017; Moffat et al., 2007). 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 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).

On the other hand, 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.

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, as revealed in the plots of Fc. 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

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020


Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

not to be used in any kind of environmental modelling related to the reviewed drivers and fluxes, whatsoever.

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 is available.

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.

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020


Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

5. Conclusions

Eight different gap-filling algorithms for estimating 16 meteorological drivers as well as 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 eight algorithms, the RF showed the highest level of robustness and reliability in estimating the Fc, as its closest rival, the XGB, could not capture the minimum values equally well, despite providing slightly better RMSE and R^2. 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 level as the rest ML algorithms and was suspected of enduring over-fitting issues. 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.

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020


Geoscientific Instrumentation Methods and Data Systems Discussions

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) 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.

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.

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.

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

754 Aubinet, M., Grelle, A., Ibrom, A., Rannik, Ü., Moncrieff, J., Foken, T., Kowalski, A. S., Martin, P. H., Berbigier, P., Bernhofer, C.,
755 Clement, R., Elbers, J., Granier, A., Grünwald, T., Morgenstern, K., Pilegaard, K., Rebmann, C., Snijders, W., Valentini, R. and
756 Vesala, T.: Estimates of the Annual Net Carbon and Water Exchange of Forests: The EUROFLUX Methodology, Adv. Ecol. Res., 30,
757 113–175, doi:10.1016/S0065-2504(08)60018-5, 1999.

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

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

761 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
762 aspen-hazelnut forest in relation to net ecosystem production, Agric. For. Meteorol., 126(3–4), 237–255,
763 doi:10.1016/J.AGRFORMET.2004.06.011, 2004.

764 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.,
765 Law, B. E., Margolis, H. A., Mccaughey, J. H., Munger, J. W., Oechel, W. and Schaeffer, K.: Use of change-point detection for friction-
766 velocity threshold evaluation in eddy-covariance studies, Agric. For. Meteorol., 171–172, 31–45, doi:10.1016/j.agrformet.2012.11.023,
767 2013.

768 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.,
769 Norton, J. P., Perrin, C., Pierce, S. A., Robson, B., Seppelt, R., Voinov, A. A., Fath, B. D. and Andreassian, V.: Characterising
770 performance of environmental models, Environ. Model. Softw., 40, 1–20, doi:10.1016/j.envsoft.2012.09.011, 2013.

771 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.,
772 Ewenz, C., Grace, P., Griebel, A., Haverd, V., Hinko-Najera, N., Huete, A., Isaac, P., Kanniah, K., Leuning, R., Liddell, M. J.,
773 MacFarlane, C., Meyer, W., Moore, C., Pendall, E., Phillips, A., Phillips, R. L., Prober, S. M., Restrepo-Coupe, N., Rutledge, S.,
774 Schroder, I., Silberstein, R., Southall, P., Sun Yee, M., Tapper, N. J., Van Gorsel, E., Vote, C., Walker, J. and Wardlaw, T.: An
775 introduction to the Australian and New Zealand flux tower network - OzFlux, Biogeosciences, 13(21), 5895–5916, doi:10.5194/bg-13-
776 5895-2016, 2016a.

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

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

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

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

787 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
788 component carbon fluxes in a semi-arid *Acacia* woodland, central Australia, J. Geophys. Res. Biogeosciences, 118(3), 1168–1185,
789 doi:10.1002/jgrg.20101, 2013.

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

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

793 Dreyfus, S. E.: Artificial neural networks, back propagation, and the kelley-bryson gradient procedure, J. Guid. Control. Dyn., 13(5),
794 926–928, doi:10.2514/3.25422, 1990.

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

797 Falge, E., Baldocchi, D., Olson, R., Anthoni, P., Aubinet, M., Bernhofer, C., Burba, G., Ceulemans, R., Clement, R., Dolman, H., Granier,
798 A., Gross, P., Grünwald, T., Hollinger, D., Jensen, N. O., Katul, G., Keronen, P., Kowalski, A., Lai, C. T., Law, B. E., Meyers, T.,
799 Moncrieff, J., Moors, E., Munger, J. W., Pilegaard, K., Rannik, Ü., Rebmann, C., Suyker, A., Tenhunen, J., Tu, K., Verma, S., Vesala,
800 T., Wilson, K. and Wofsy, S.: Gap filling strategies for defensible annual sums of net ecosystem exchange, Agric. For. Meteorol.,
801 107(1), 43–69, doi:10.1016/S0168-1923(00)00225-2, 2001.

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020

Geoscientific Instrumentation Methods and Data Systems Discussions

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.

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

847 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
848 covariance methane fluxes: A comparison of three machine learning algorithms and a traditional method with principal component
849 analysis, Glob. Chang. Biol., 26(3), 1499–1518, doi:10.1111/gcb.14845, 2020.

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

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

854 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.,
855 Granier, A., Grelle, A., Hollinger, D., Janssens, I. A., Jarvis, P., Jensen, N. O., Katul, G., Mahli, Y., Matteucci, G., Meyers, T., Monson,
856 R., Munger, W., Oechel, W., Olson, R., Pilegaard, K., Paw U H, K. T., Thorgeirsson, H., Valentini, R., Verma, S., Vesala, T., Wilson, K.
857 and Wofsy, S.: Jourassess2, Agric. For. Meteorol., 113(113), 97–120, 2002.

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

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

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

864 Moffat, A. M., Papale, D., Reichstein, M., Hollinger, D. Y., Richardson, A. D., Barr, A. G., Beckstein, C., Braswell, B. H., Churkina, G.,
865 Desai, A. R., Falge, E., Gove, J. H., Heimann, M., Hui, D., Jarvis, A. J., Kattge, J., Noormets, A. and Stauch, V. J.: Comprehensive
866 comparison of gap-filling techniques for eddy covariance net carbon fluxes, Agric. For. Meteorol., 147(3–4), 209–232,
867 doi:10.1016/j.agrformet.2007.08.011, 2007.

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

870 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
871 ecological memory in plant and ecosystem processes, Ecol. Lett., 18(3), 221–235, doi:10.1111/ele.12399, 2015.

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

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

876 Reichle, R. H., Koster, R. D., Dong, J. and Berg, A. A.: Global soil moisture from satellite observations, land surface models, and
877 ground data: Implications for data assimilation, J. Hydrometeorol., 5(3), 430–442, doi:10.1175/1525-
878 7541(2004)005<0430:GSMFSO>2.0.CO;2, 2004.

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

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

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

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

888 Scanlon, T. M. and Kustas, W. P.: Partitioning carbon dioxide and water vapor fluxes using correlation analysis, Agric. For. Meteorol.,
889 150(1), 89–99, doi:10.1016/j.agrformet.2009.09.005, 2010.

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

https://doi.org/10.5194/gi-2020-21
Preprint. Discussion started: 7 September 2020

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

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

894 Tannenbaum, C. E.: The empirical nature and statistical treatment of missing data., Diss. Abstr. Int. Sect. A Humanit. Soc. Sci., 70(10-
895 A), 3825 [online] Available from: http://gateway.proquest.com/openurl?url_ver=Z39.88-
896 2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&res_dat=xri:pqdiss&rft_dat=xri:pqdiss:3381876%5Cnhttp://ovidsp.ovid.com/ovi
897 dweb.cgi?T=JS&PAGE=reference&D=psyc7&NEWS=N&AN=2010-99071-044, 2010.

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

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

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

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

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

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

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

910