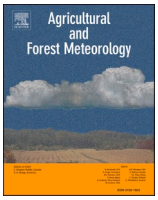
Agricultural and Forest Meteorology 325 (2022) 109114

Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

journal homepage: www.elsevier.com/locate/agrformet



Multiple gap-filling for eddy covariance datasets

Antje M. Lucas-Moffat [a,b,*], Frederik Schrader [a], Mathias Herbst [b], Christian Brümmer [a]

[a] Thünen Institute of Climate-Smart Agriculture, Bundesallee 65, 38116 Braunschweig, Germany
[b] German Meteorological Service (DWD), Centre for Agrometeorological Research, Bundesallee 33, 38116 Braunschweig, Germany

ARTICLE INFO

Keywords:
Eddy covariance fluxes
trace gases
artificial gap scenarios
bootstrapping
multiple gap-filling
ensemble results

ABSTRACT

With novel developments in technology, eddy covariance flux measurements have become feasible for a variety of trace gases. While the statistical properties and gap-filling strategies have been well examined for carbon dioxide, these are much less understood for other gases.

Here, we propose a universal methodology deploying multiple gap-filling techniques and artificial gap scenarios to evaluate the techniques' performances, infer the statistical flux properties, and fill the real gaps in eddy covariance datasets of any trace gas. The methodology was implemented in a gap-filling framework with techniques spanning from simple and diurnal interpolations, look-up tables, artificial neural networks, to an inferential model. For the new scheme of half-hourly and daily artificial gaps, each additional gap was superimposed one at a time (thus keeping the disturbance to a minimum) for the whole dataset and the scenarios were resampled by bootstrapping. The gap-filled sums were then estimated from the ensemble of well-performing gap-filling techniques. The gap-filling framework was applied to campaign data of three different trace gases (51 days of ammonia, 79 days of total reactive nitrogen, and 89 days of methane flux measurements). The aggregated fluxes are stated as ensemble ranges of multiple techniques plus the techniques' uncertainties. Additionally, the framework was used to gap-fill a full year of carbon dioxide flux measurements yielding similar performances as previously reported.

Based on a review of gap-filling comparison studies and on our findings, we suggest reconsidering the standard procedure of using *one* gap-filling technique for multi-site studies. Deploying *multiple* gap-filling techniques and providing *ensemble* results of gap-filled sums will help to minimize the influence of a single technique and thus lead to a more robust flux aggregation. Furthermore, the estimated overall uncertainty will be more realistic by accounting for the ensemble range of multiple techniques.

1. Introduction

Eddy covariance has become the preferred method for continuous long-term monitoring of carbon dioxide (CO_2) and energy exchange between terrestrial ecosystems and the atmosphere and allows for an assessment of ecosystem metabolism over time scales from hours to decades (Baldocchi, 2019; Odum, 1969; Vernadsky, 1998). The integration of eddy covariance datasets with biometeorological drivers enables researchers to investigate how plants respond to a number of environmental and biological forcings such as light, temperature, water availability, and phenology (Brümmer et al., 2012; Brümmer et al., 2008; Keenan et al., 2014; van Dijk et al., 2005) as well as how the metabolism of whole ecosystems is responding to longer term trends in the environment; these include changes in atmospheric nitrogen deposition (Fernández-Martínez et al., 2014; Fleischer et al., 2013; Magnani et al., 2007), rising CO_2 concentration and temperature (Keenan et al., 2013), or rather episodic disturbances such as wind throws (Lindauer et al., 2014), insect infestation (Brown et al., 2010), and heat waves (Graf et al., 2020).

A global network of long-term CO_2 and water vapor flux measurements has existed since the late 1990s (Baldocchi et al., 2001) and flux towers are nowadays organized in continental-scale research infrastructures like the Integrated Carbon Observation System (ICOS) in Europe (Heiskanen et al., 2022) or the National Ecological Observation Network (NEON) in the US (Metzger et al., 2019) using standardized processing tools (Pastorello et al., 2020). However, continuous eddy covariance measurements of other trace gases like methane (CH_4) or nitrous oxide (N_2O) and air pollutants like ammonia (NH_3) or nitrogen oxides (NO_x) have largely remained experimental due to the technical complexity and large equipment and operational costs involved

* Corresponding author:
E-mail address: amm@moffats.de (A.M. Lucas-Moffat).

https://doi.org/10.1016/j.agrformet.2022.109114
Received 21 March 2022; Received in revised form 31 July 2022; Accepted 1 August 2022
Available online 12 September 2022

(Flechard et al., 2011).

Recent technological advancements in the use of tunable diode laser absorption spectrometers (TDLAS) and quantum cascade lasers (QCL) resulted in accurate analytical devices with high precisions and fast response times, enabling eddy covariance measurements of field-scale N_2O and CH_4 fluxes (Denmead et al., 2010; Neftel et al., 2010; Tang et al., 2018). A steadily rising number of campaigns using harmonized setups and data processing (Nemitz et al., 2018) offers the opportunity to synthesize findings of Non-CO_2 greenhouse gases (Knox et al., 2019) to improve the understanding of land surface-atmosphere interactions and ecosystem functioning in a broader context.

Eddy covariance flux measurements of NH_3 (Famulari et al., 2004; Sintermann et al., 2011; Zöll et al., 2016) and other reactive nitrogen (N_r) compounds (Ammann et al., 2012; Brümmer et al., 2013; Wintjen et al., 2022) have been extremely limited in numbers and are subject to substantial uncertainty due to challenging fast-response detection (Marx et al., 2012), issues regarding inlet design, sampling losses, and air column chemical reactions for highly reactive and soluble N_r species (Horii et al., 2006; Horii et al., 2004). With recent progress in accounting for high-frequency losses (Wintjen et al., 2020) and highlighting opportunities for a more robust atmospheric deposition monitoring and modeling (Schrader et al., 2020), a closer coupling between research dealing with inert greenhouse gases and reactive nitrogen species is expected.

Regardless of the species of interest, all eddy covariance measurements are subject to data gaps of various lengths. Gaps arise in all situations where the basic assumptions of eddy covariance theory – such as fully turbulent fluxes, a uniform and horizontal footprint, the average of fluctuations equaling zero, and negligible density fluctuations (Aubinet et al., 2012) – are violated. Other reasons for missing data may be instrument failures or implausible spikes in the data of unknown origin. The gaps in the eddy covariance datasets need to be filled in order to calculate sums of the fluxes over certain time periods. Reliable sums are required to provide accurate estimates of whole ecosystem greenhouse gas budgets, which form the basis for assessing land management practices and for developing climate-related conservation guidelines.

For CO_2, a multitude of gap-filling techniques has been developed with a first comparison of three methods in Falge et al. (2001), fifteen different techniques in Moffat et al. (2007), and nine techniques in a newer study by Mahabbati et al. (2021). With Non-CO_2 eddy covariance measurements coming up and advances in computer science, the use of different machine learning techniques has been explored predominantly for CH_4 (Irvin et al., 2021; Kim et al., 2019). Even for CO_2, new gap-filling techniques are still being developed, usually not aiming purely at gap-filling the fluxes but also providing robust uncertainty estimates (e.g. Menzer et al., 2013; Vitale et al., 2019b; Wang et al., 2015). The uncertainty of the measured eddy covariance fluxes is largely due to a random measurement error (Hollinger and Richardson, 2005). This error can be determined from the residuals of the gap-filled fluxes (Richardson et al., 2008). An additional source of uncertainty is induced by the gap-filling itself.

Gap-filling means reconstructing the missing flux measurements in the time series. A gap-filling technique makes use of the information contained in the existing flux and ancillary measurements and the available techniques cover a wide range of methods such as interpolation, parameterization of semi-empirical equations, modeling of underlying processes, and machine learning. The main source of uncertainty and critical error for the sums of the gap-filled flux time series is the systematic error. Even small but systematic offsets in the gap-filled fluxes will add up linearly over the aggregated time period. The gap-filling performance of a single technique can be estimated by inserting additional artificial gaps (e.g. Falge et al., 2001; Moffat et al., 2007; Richardson and Hollinger, 2007) or statistical inference (e.g. Menzer et al., 2013; Vitale et al., 2019b; Wang et al., 2015). However, the error bounds found for the techniques often underestimate the differences found between the gap-filled annual sums, see Appendix A.2.1. The reason is that the gap-filling error is not only dependent on the general performance of the gap-filling technique but highly influenced by the specific characteristics of fluxes and real gaps in the dataset: the site properties (e.g. climate, footprint, sensors, and setup), the ecosystem (e.g. type, species, soil, seasonality, management), the flux characteristics (e.g. magnitude of the flux, noise level, diurnality, gas concentration, storage), the availability of ancillary data (e.g. measurements of air and soil meteorology, radiation, vegetation parameters), and the type of gaps (e.g. position, length, amount). Some of these factors vary not only from site to site but also from year to year. Besides, though the gap-filling techniques interpolate the fluxes in time, reconstructing the missing data may require extrapolating to conditions not represented in the measured data used for the gap-filling. Since each technique will be differently affected by the specific dataset characteristics and the degree of extrapolation, the *choice of the gap-filling technique* has a large impact on the annual sum estimates. Falge et al. (2001) already emphasized the need to standardize gap-filling methods to improve the comparability of flux data products across sites.

To improve the comparability of the gap-filled sums, one approach is to use the most suitable gap-filling technique for each dataset. A technique geared specifically to the dataset characteristics mentioned above usually has the lowest error in absolute terms and hence yields very reliable sums. This is the main reason for many site PIs to develop their own gap-filling routines and also for the development of new techniques in general. Usually meta-comparisons across single site papers are based on sums obtained with different gap-filling techniques.

Since this procedure is not feasible for large multi-site studies, the common practice for inter-site comparisons has been the approach to choose *one gap-filling technique* with an overall good performance and ease of implementation (e.g. in the two FLUXNET datasets Drought, 2018; Pastorello et al., 2020). However, the comparability of the estimated sums is only improved if using one gap-filling technique reduces the relative error of the gap-filled annual sums between sites.

Revisiting reported annuals sum estimates of CO_2 of Moffat et al. (2007) in Appendix A.2.2 shows that the mean difference in annual sums for the same technique between sites is very similar to the mean difference for the same site between techniques. The same gap-filling technique might even have a large underestimate one year and a large overestimate the next year at the same site. As far as we know, the potential advantage that using one technique reduces the relative errors has not been proven.

Besides, even if the relative errors were reduced, the absolute values of the gap-filled fluxes and annual sums are usually used in the further analysis and discussed in the results. The reported uncertainty of the gap-filled sums needs to properly account for the gap-filling uncertainty. A more generic gap-filling error can only be quantified using a wide range of techniques and datasets. For CO_2, Moffat et al. (2007) only quantified this error on a small subset of sites of six forested ecosystems. The effect of the position of longer gaps has been investigated on the same subset of sites but only for one gap-filling technique (Richardson and Hollinger, 2007). For CH_4, Kim et al. (2019) quantified the gap-filling error spanning multiple techniques for each of the site years. Generally speaking, this generic gap-filling error needs to be quantified for a wider scope of sites differing in dataset characteristics or even better specifically for each dataset. Moreover, for a lot of the trace gases, the suitability of different gap-filling techniques has to be evaluated in the first place.

To address these needs, we suggest that rather than to use only one technique for gap-filling, an *ensemble of gap-filling techniques* should be implemented. For this, we propose a methodology of a universal gap-filling framework which can be commonly applied to any kind of eddy covariance dataset. In this manuscript, the methodology has been implemented deploying different types of gap-filling techniques, bootstrapping artificial gap scenarios, and using model residuals for the statistical analysis. This allows evaluating the performance of the techniques, quantifying the generic gap-filling error for each dataset, and

reporting ensemble sums plus uncertainties. The usefulness and universal applicability of such a gap-filling framework will be demonstrated on four eddy covariance datasets of different trace gases. The multiple gap-filling tool and code developed for this manuscript are publicly available at https://doi.org/10.18160/R6S5-47J1.

2. Methods and materials

2.1. Methodology

Here, we propose a new methodology to gap-fill fluxes and estimate the uncertainty in any kind of eddy covariance dataset[1]. The term methodology is used since it describes a universal framework with an interchangeable set of specific methods. The steps of the methodology are as follows:

1 **Choice of data:** Select the subset of half-hourly eddy fluxes with very high quality according to established quality control methods (e.g. Mauder and Foken, 2006) and flag the other data points in the time series as gaps. If discontinuities during the measurement period exist (such as significant modifications in the experimental setup, major changes in the ecosystem state e.g. due to harvest or fire, substantial gaps of more than two weeks), it may be advisable to split the dataset and fill the subsets prior and subsequent to the event separately.
2 **Suite of gap-filling techniques:** To encompass the effect of the choice of the gap-filling technique on the specific dataset characteristics, choose a suite of techniques of different type and complexity like simple interpolations, look-up tables, regression methods, machine learning algorithms, or mechanistic models.
3 **Artificial gap scenarios:** Generate multiple scenarios of artificial gaps and fill these with the suite of gap-filling techniques.
4 **Statistical properties:** Analyze the model residuals, i.e. the difference between the observed and predicted data points, to evaluate the performance of the gap-filling techniques and to infer the statistical properties of the eddy covariance dataset. The evaluation of the gap-filling technique depends on the purpose of the gap-filled dataset. For aggregating fluxes, the basic requirement of a gap filling technique is a small bias error centered around zero.
5 **Multiple gap-filling of real gaps:** Use all gap-filling techniques with acceptable performance, i.e. fulfilling the basic requirements, to fill the *real* gaps in the dataset. The fluxes are aggregated over the designated periods for each technique.
6 **Ensemble results:** Report the results of the gap-filling ensemble, i.e. of the multiple sums, plus the uncertainties.
7 **Scope of the results:** The obtained results are trace gas and dataset specific and depend on the properties of each site (e.g. climate, footprint, sensors), the ecosystem (e.g. type, seasonality, management), the flux characteristics (e.g. magnitude, noise level, diurnality), the availability of ancillary data (e.g. meteorology, gas concentration, ecosystem state), and the real gaps (e.g. position, length, amount).

This manuscript describes an implementation of this methodology and its application to eddy flux datasets of four different trace gases.

2.2. Dataset and site descriptions

Eddy covariance flux datasets from measurement campaigns of three different trace gases were chosen as examples of the proposed methodology: ammonia (NH_3), total reactive nitrogen (tN_r), and methane (CH_4). Additionally, one annual dataset of carbon dioxide (CO_2) measurements from Moffat et al. (2007) was included for comparison with the former gap-filling results. Each dataset was set up and processed specific to its properties in the multiple gap-filling tool (see Appendix A.1). An overview with respect to their gap-filling properties and settings is provided in Table 1 and details on the distributions of gaps can be found in Fig. 2, 6, 10, and 14. The few short gaps in the ancillary micrometeorological and concentration measurements were pre-filled by linear interpolation. In cases where the datasets were separated into daytime and nighttime data, the threshold of the global radiation for daytime was set to 5 W m^{-2}.

2.2.1. NH_3 – Bourtanger Moor

Turbulent exchange fluxes of ammonia (NH_3) were measured above an ombrotrophic peatland in Northwestern Germany ("Bourtanger Moor") at 52°39′21″ N and 7°11′00″ E from February to May 2014 using the eddy covariance technique. Vegetation at the site mainly consisted of bog heather (*Erica tetralix*), purple moor-grass (*Molinia caerulea*), cotton grass (*Eriophorum vaginatum, E. angustifolium*), and a few birches (*Betula pubescens*) and Scots pines (*Pinus sylvestris*). Further site details are given in Hurkuck et al. (2014) and Hurkuck et al. (2016).

Fast response (10 Hz) NH_3 mixing ratios were measured with a quantum cascade laser (Mini QC-TILDAS-76, Aerodyne Research, Inc, Billerica, MA. USA) and vertical wind velocities with a 3D sonic anemometer (R3-50, Gill Instruments, Lymington, UK). A detailed description of the measurement setup and data processing steps is given in Zöll et al. (2016).

2.2.2. tN_r – Bavarian Forest

The eddy covariance measurements of total reactive nitrogen (tN_r) were taken above a mixed forest stand in the Bavarian Forest National Park, Germany, at 48°56′33″ N and 13°25′11″ E from July to September 2016. Vegetation at the site was dominated by spruce (*Picea abies*, ~80% of the flux footprint area) and beech (*Fagus sylvatica*, ~20%).

A TRANC system (Total Reactive Atmospheric Nitrogen Converter; Marx et al., 2012) at 30 m above ground at 10 Hz frequency using a custom-built converter system to nitrous oxide in conjunction with a chemiluminescence detector (CLD 780 TR, ECO PHYSICS AG, Dürnten, CH) were used to measure tN_r concentrations at a frequency of 10 Hz. The TRANC converts all reactive nitrogen compounds to nitrogen monoxide (NO), which is finally analyzed in the CLD. The fast time response of both converter and analyzer and high conversion efficiency of the TRANC allowed for eddy-covariance flux estimations in combination with a 3D sonic anemometer (R3-50, Gill Instruments, Lymington, UK) for the vertical wind velocities. Additional information on the site and measurements is given in Zöll et al. (2019) and Wintjen et al. (2022).

2.2.3. CH_4 – Skjern Wetland

The methane (CH_4) fluxes were measured from March to May 2010 at a site located on a floodplain close the mouth of the Skjern River in Western Denmark at 55°54′46″ N and 8°24′17″ E. The footprint area of the micrometeorological measurements is almost entirely covered by a restored wetland consisting of meadows, wetlands, lakes, and meandering water courses. The meadows are managed by both grazing and hay making.

Eddy flux measurements of CH_4 have been conducted at the Skjern site since 2008 at 7 m height with a 3D sonic anemometer (R3-50, Gill Instruments, UK) and a DLT-100 gas analyzer (Los Gatos Research Inc., Mountain View, CA, USA) which is based on the 'off-axis integrated cavity output spectroscopy' technology (OA-ICOS). Further details about the site, vegetation, and instrumentation can be found in Herbst et al. (2011).

2.2.4. CO_2 – Hainich Forest

The eddy covariance measurements of carbon dioxide (CO_2) were taken above a deciduous broadleaf forest at Hainich, Germany, in the

[1] The basic principles of the methodology may be applied also to flux datasets obtained with other technologies such as the aerodynamic gradient method for reactive gases.

year 2000. The same dataset was used in the gap-filling comparison by Moffat et al. (2007).

The Hainich forest is dominated by beech (*Fagus sylvatica*, 65%), ash (*Fraxinus excelsior*, 25%), and maple (*Acer pseudoplantanus* and *Acer plantanoides*, 7%) with the flux tower located at 51°04′46″ N and 10°27′08″ E.

The turbulent exchange of CO_2 is measured above the canopy at 43.5 m with a 3D sonic anemometer (R3-50, Gill Instruments, Lymington, UK) and a LI-6262 infrared gas analyzer (LICOR, Lincoln, Nebraska, USA) located at the base of the tower. Further details on the instrumentation and vegetation can be found in Knohl et al. (2003) and Kutsch et al. (2008).

2.3. Gap-filling techniques

Gap-filling techniques make use of the obvious and hidden relationships between the flux measurements and the ancillary data. One key requirement of the proposed methodology is to use different types of gap-filling techniques to exploit different characteristics of the datasets. For our analysis, we used interpolation methods, look-up tables, artificial neural networks, and an inferential model. A summary with abbreviations of the gap-filling techniques used herein can be found in Table 2.

2.3.1. Simple interpolation methods

One of the simplest methods for gap-filling is linear interpolation (*IP_lin*) between the previous and next existing data point. For small gaps ($\leq$12 half-hours), the gaps are linearly interpolated. For longer gaps, a pure linear interpolation between half-hourly data points may lead to very unrealistic estimates and offsets. Therefore, longer gaps are filled by linear interpolation between daily means.

Another simple interpolation is the moving average (*IP_mov*). The gaps are filled by averaging (rolling) over a window, here five half-hours at a time. If there are not enough data points (n<2) within the five half-hours, the window size of adjacent half-hours is increased in steps of two (5, 7, 9, 11, …) and *IP_mov* reapplied for the whole dataset.

The simple interpolation methods have the advantage that any measurement series can be filled independently of ancillary information. The time series is treated as pure consecutive half-hours hh_i (with index i = 1, …, N where N is the total number of data points). The only underlying assumption is that there is some relationship between the consecutive data points, which in a first order approximation is linear.

2.3.2. Diurnal interpolation methods

The main difference between the simple interpolation methods and the diurnal interpolation methods is that the information of the *time of day* is included. The time steps are indexed for the day d (with $d = 1, \ldots, N_d$ where N_d is the total number of days) and time of day t (with $t = 1,\ldots,$ 48 for each half-hour), i.e. $hh_{d,t}$. To fill larger gaps, the window size is not extended to adjacent half-hours but rather to half-hours at the same time of the previous and next day(s).

The weighted daylight mean (*WDM*) was used in the analysis of the gap-filling comparison for estimating daily sums from incomplete data (Moffat et al., 2007) and will be tested here as a gap-filling technique. The gaps are filled with the mean of the daylight fluxes during daytime and with the mean of the nighttime fluxes during nighttime. The window size starts with the current day. If there are not enough data points (n<2) in the current day, the window is increased to adjacent days in steps of ±1 day.

To reduce the measurement noise, the Non-CO_2 fluxes can be averaged over fixed time periods. Here, a fixed diurnal average (*FDA_6hh*) of three hours (0:00–03:00, …, 21:00–24:00) will be used for filling the gaps with the mean of these six half-hour intervals. The window size starts with the current day. If there are not enough data points (n<2) in the current day, the window is increased to adjacent days in steps of ±1 day.

For the moving diurnal average (*MDA_5hh*), the gaps are first filled with the moving average of ±1 hour, i.e. 5 half-hours in total. The window size starts with the current day and in this case is the same as *IP_mov* described above. If there are not enough data points (n<2) in the current day, the window size is increased to adjacent days in steps of ±1 day. *MDA* is part of the marginal distribution sampling (*MDS*) (Reichstein et al., 2005) which combines this technique with a look-up table (see next section).

The mean diurnal course (*MDC*_d3) after (Falge et al., 2001) considers solely the current half-hour and starts with a window size of ±3 days. If there are not enough data points (n<2) in these seven half-hours, the window is increased in steps of ±3 days. Additionally, the same algorithm is performed with a window size of ±7 days (*MDC*_d7).

All these methods take advantage of the fact that the observed net fluxes are the result of biological processes that often exhibit a diurnal course. These techniques are based on the time stamp and can be used without any ancillary measurements. Only *WDM* additionally needs the information of daylight; though if no radiative measurements are available, the potential radiation calculated from latitude and longitude can also be used.

Essentially, the diurnal interpolation methods are simple look-up tables with the fluxes binned to certain times of day(s).

2.3.3. Look-up tables

A more sophisticated approach to "look-up" the values is by binning the fluxes depending on correlated variables (Falge et al., 2001; Reichstein et al., 2005). For the look-up table (LUT) with one independent variable *V1*, the gaps are filled with the mean (i.e. bin-average) of all fluxes within a certain range of *V1* and a window size of ±3 days (*LUT_V1_d3*). If there are not enough data points (n<2) in these seven days, the window size is increased in steps of ±3 days. Additionally, the same algorithm is performed with a window size of ±7 days (*LUT_V1_7d*). As a refinement of the look-up, the number of independent variables is increased to two (*LUT_V1V2_3d & LUT_V1V2_7d*) and then three (*LUT_V1V2V3_3d & LUT_V1V2V3_7d*).

The look-up tables exploit the correlations between the fluxes and the ancillary measurements of gas concentrations and climatic conditions. The best choices of the independent variables are the ones with highest correlation and only little gaps. For carbon fluxes, light, temperature, and humidity are sufficient variables since these are the main drivers of ecosystem respiration and photosynthesis. For Non-CO_2 gases, the main drivers of the fluxes may not be known and other variables like concentration might be correlated but not necessarily drivers. Here, the variables with highest (nonlinear) correlations were pre-determined for each dataset using the ANN approach described in Moffat et al. (2010).[2] The choice of input variables for the four datasets is provided in Table 1. Using more dependent variables for the look-up table means more similar conditions during the flux measurements but less available data points for filling the gap. The bin width is also a trade-off between the similarity of the conditions and the availability of flux measurements. Analog to typical margins used for CO_2, (e.g. in the *MDS* algorithm described below), the total range of a variable divided by roughly sixteen was taken as the bin width.

The *MDS* algorithm (Reichstein et al., 2005; Wutzler et al., 2018) is a gap-filling scheme based on two gap-filling techniques, *LUT* and *MDA_5hh* (Section 2.3.2). The first two steps of the *MDS* sampling scheme are a look-up table with three independent variables

[2] The principles for finding input variables with the highest correlations described in Moffat et al. (2010) based on ANNs can be adopted using this gap-filling framework based on LUT: The higher the correlation between variables and fluxes, the higher the R^2 gap-filling performance (Eq. 2) of the LUT. By systematically using different sets of depending variables and determining the LUT R^2 performance, the input variables yielding the highest correlations may be identified.

(*LUT_V1V2V3_d7*) with first ±7 and then ±14 days. This is followed by steps of *LUT_V1_d7*, several *MDA_5hh* and again *LUT_V1V2V3_d7* for pre-defined time windows (of up to ±210 days). *MDS* is used as a standard gap-filling technique in FLUXNET (e.g. Drought, 2018; Pastorello et al., 2020) and in the freely available REddyProc tool. For comparison with *MDS*, the same *LUT* variables and ranges (Rg ±50, Ta ±2.5, VPD ±5.0)[3] as in the online REddyProc tool[4] were also applied to each dataset in *LUT_MDS*.

The *MDS* algorithm is very flexible in term of missing data not only in the fluxes but also in the three independent ancillary variables. However, if these variables are complete, then the gaps in the fluxes are mainly filled during the first step of the *MDS* algorithm plus a few more during the second step, i.e. with a look-up table *LUT_V1V2V3_d7*. Since for the four trace gas datasets the ancillary variables have been pre-filled, almost all artificial and real gaps were filled during the first (>96%) and then second (>98.5%) step. Hence, applying *LUT_MDS* very closely resembles applying the full *MDS* algorithm for the datasets presented here.

For some techniques, a few (longer) gaps could not be filled. These half-hours were excluded from the performance statistics to compare the same subset of data points. Generally, to optimize the gap-filling performance and to overcome problems like unfilled gaps, the different techniques can be combined as in the REddyProc tool. However, since for this study, our goal was to test their applicability in the first place, each gap-filling technique was applied individually. Only in the end, when calculating the total sums, the still missing half-hours were filled with *MDA_hh5* as the default technique.

2.3.4. Artificial neural networks

An artificial neural network (ANN) is a purely empirical nonlinear regression model and can also be considered a look-up table with multiple independent variables and continuous bins but one window for the full dataset. The ANN algorithm used is based on the classical back-propagation algorithm. The training of each network is performed by propagating the input data through the nodes via the weighted connections and then back-propagating the error and adjusting the weights so that the ANN output optimally approximates the fluxes. Each ANN training was repeated ten times and the modeled results were averaged. Details on the training algorithm and the C++ framework can be found in Moffat (2012).

For a better comparison with LUT, the ANNs were first trained with the same three variables as used for *LUT_V1V2V3* (see Table 1). Additionally, the ANNs were trained with all available input variables (*ANN_all*). (No additional inputs like fuzzies for the season were generated, since the three Non-CO_2 datasets were spanning only two to three months.)

After having been trained on a specific dataset, an artificial neural network maps the underlying dependencies of the fluxes from the provided input variables. No prior knowledge is required for these self-learning algorithms but ANNs have the disadvantage that complete and evenly scalable input data is required. Other machine learning algorithms could be added to the suite of gap-filling techniques, e.g. support vector machines or random forests. The latter are more flexible in dealing with gaps or outliers in the input data and have been shown to match or even outperform ANNs on CH_4 (e.g. Irvin et al., 2021).

2.3.5. Inferential model

Complex models can also be used for gap-filling biosphere-atmosphere fluxes. As an example, we included modeled fluxes of NH_3 using an inferential model after Nemitz et al. (2001). For the unmanaged site Bourtanger Moor, the ground-layer of the two-layer canopy compensation point model was switched off and the model was run in a one-layer configuration. The average annual total (wet and dry) N input, as a driving parameter for the stomatal emission potential, was estimated to be around 25 kg N ha^{-1} yr^{-1} (Hurkuck et al., 2014). The aerodynamic, quasi-laminar, and cuticular resistances were parameterized as described in Massad et al. (2010) for "semi-natural/moorland ecosystem type", and the stomatal resistance was parameterized after Wesely (1989). The modeled NH_3 fluxes at Bourtanger Moor are further discussed and compared to measurements in Zöll et al. (2016). Meteorological drivers (air temperature, relative humidity, shortwave radiation, atmospheric pressure, friction velocity, and Obukhov length) and NH_3 concentrations necessary to run the model were measured at the site and averaged to half-hourly values as described in detail by Zöll et al (2016). Relative humidity was estimated from IRGA measurements due to a malfunction of the dedicated relative humidity sensor. Wind speed at the reference height was calculated from the friction velocity and the logarithmic wind profile for internal consistency of the model. Air temperature and relative humidity at the notional mean height of trace gas exchange were extrapolated from measured values at the reference height and their respective turbulent fluxes as described in Nemitz et al., 2009. Seasonal averages of leaf area index (*LAI*) and canopy height were taken from Table 6 in Massad et al. (2010).

As the inferential model (*Model_NH3)* was parameterized independently of the NH_3 flux data measured at the Bourtanger Moor site, the artificial gap scenarios (see Section 2.4) would not influence any modelled results. Hence, these could be taken directly as the secondary dataset for the 'hhs' and 'days' scenarios.

More types of gap-filling techniques could potentially be added to the analysis. One big group are nonlinear regression methods based on semi-empirical equations. However, these are highly trace gas specific (e.g. equations of respiration and photosynthesis for CO_2). For other trace gases with sporadic flux bursts such as N_2O, filling gaps in the datasets remains challenging (Nemitz et al., 2018). The simple and diurnal interpolation methods have the advantage that these require no additional input data as the availability of ancillary measurements in campaigns might be limited or the driving processes of the fluxes might be unknown or not directly measurable.

2.4. Artificial gap scenarios

To be able to compare gap-filled (predicted) with measured (observed) fluxes, artificial gaps are superimposed on the datasets which already have real gaps in their measurements. In Moffat et al. (2007), the position of ten percent artificial gaps were prescribed in five different artificial gap length scenarios with ten permutations each. Despite the ten permutations, the position of the additional gaps still influenced that analysis and only ~65%[5] of the data were sampled.

Here, we used a new procedure for superimposing the artificial gaps. To keep the influence of the additional artificial gaps as small as possible, only *one artificial gap at a time* is superimposed on the dataset and then gap-filled. Starting at the first data-point, the next artificial gap is placed adjacent to the previous one until the whole dataset has been scanned. The result is a secondary dataset where each data point in the time series (including the real gaps) has been gap-filled (Fig. 1). This way, rather than prescribing the position of the artificial gaps, *all* data is sampled and replaced with artificial gaps, i.e. 100% of the available data.

Two lengths of artificial gaps were chosen: single half-hours ('hhs') and single days ('days'). For 'hhs', each half-hour individually was set to be an artificial gap and filled. For 'days', each day individually set to be an artificial gap and filled (i.e. all the half-hours of one day). This

[3] For NH_3, relative humidity was not available. Therefore, only Rg and Ta were used.

[4] The R package of REddyProc provides the option to also choose other variables and ranges.

[5] The coverage of sampling 10 times 10% of the data with replacement leads to a binomial distribution of: $1-(1-10\%)^{10}=0.65$

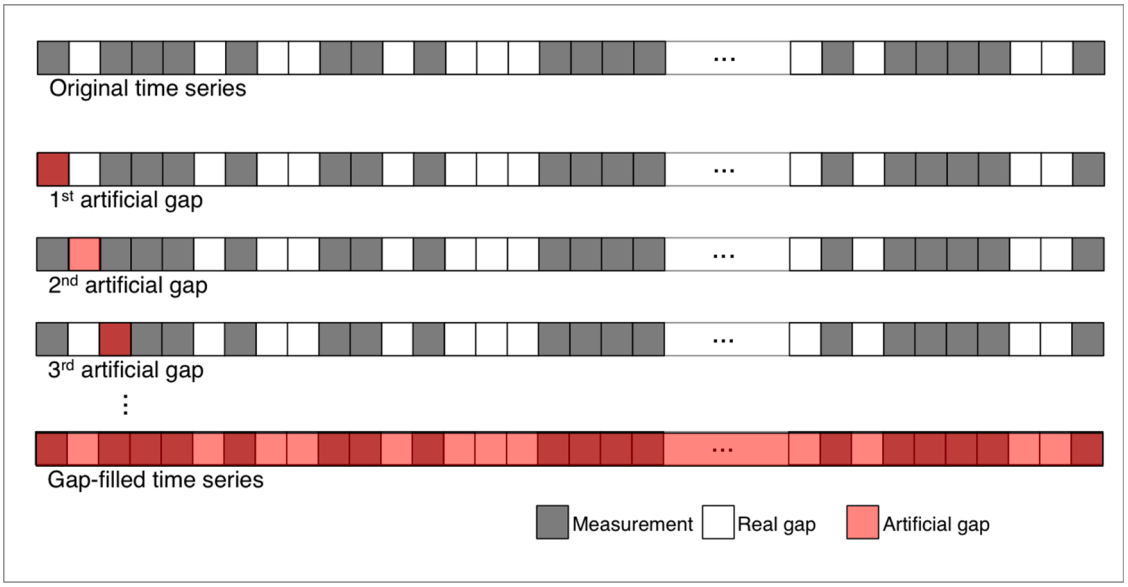


Fig. 1. Artificial gap-filling scheme.

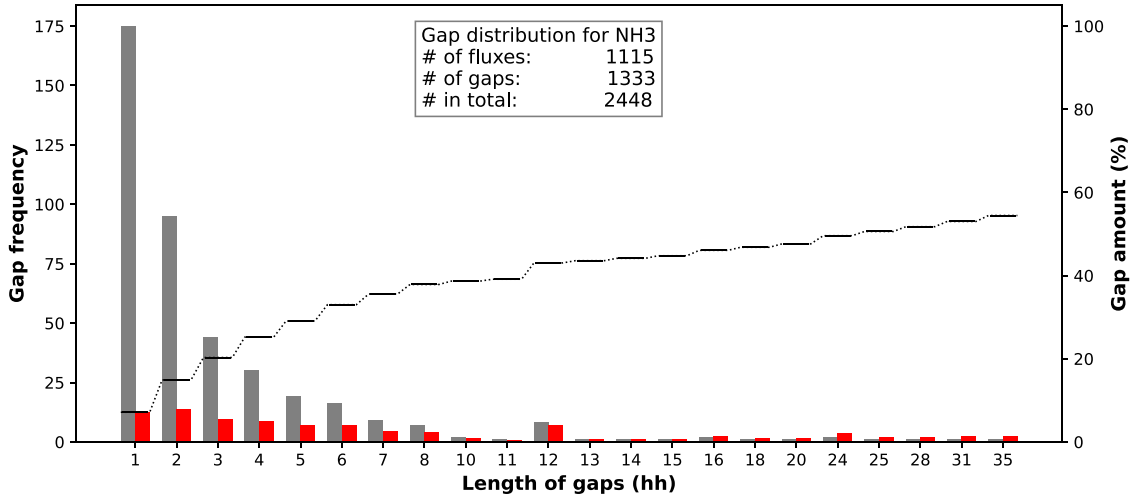


Fig. 2. Gap distribution for NH_3 sorted by gap length. The histogram denotes the frequency of occurring gaps (gray bars), their data percentage (red bars), and the accumulated total gap percentage (dotted line).

resulted in two secondary datasets (one for 'hhs' and one for 'days') for each of the gap-filling techniques. The filling of real gaps is included in the 'hhs'-scenario since either an artificial gap or a real gap is filled with no additional artificial gaps super-imposed.

The maximum artificial gap length was chosen to be a single day since this framework was primarily developed for campaign datasets which are typically short (weeks to months) and mostly continuous (few long gaps). To fill annual datasets, longer gap lengths of several days may be added to the artificial gap-filling scheme, again super-imposing one artificial gap at a time and moving through the dataset in one day steps to include the bias of the exact placement of the longer gap in the analysis.

This scheme of superimposing only one artificial gap at a time is easy to implement for all gap-filling techniques where the algorithm is centered on the gap and filled with the surrounding information or for certain time windows. For algorithms based on the whole dataset for parameterization like ANNs (Section 2.3.4), re-training might be required each time.

Here, for the 'days'-scenarios, new ANNs were trained for each scenario with one day removed (strapped). As an example, for the CH_4 dataset with 89 days of data times the ten training repetitions, this resulted in 890 trained ANNs. However, for the 'hhs'-scenarios, the results modeled from ANNs trained on the full dataset were taken since dropping a single half-hour would not have changed the modeled results. Each repetition of the training results in a different network and slightly different outputs. These differences are larger than the effect of dropping one out of over 1000 data points.

2.5. Statistical properties

2.5.1. Performance metrics

The performance of the gap-filling techniques and the uncertainties of the fluxes can be estimated from the model residuals $(p_j - o_j)$ where p_j is the predicted flux and o_j the observed flux of each half-hour j. An important metric of performance for gap-filling is the systematic error (bias error BE) summed over the number of predicted half-hourly fluxes N_p:

$$BE = \frac{1}{N_p}\sum_{N_p}(p_j - o_j). \quad (1)$$

The coefficient of determination (R^2) describes how much of the variance can be predicted by the techniques:

$$R^2 = \frac{\{\sum(p_j - \overline{p})(o_j - \overline{o})\}^2}{\sum(p_j - \overline{p})^2 \sum(o_j - \overline{o})^2} \quad (2)$$

with overbars denoting the arithmetic mean of a variable.

Since the probability density function of the flux residuals follows a Laplace distribution rather than a Gaussian (for details see Richardson et al., 2012; Vitale et al., 2019a), the standard deviation (*SDev*) was calculated from the mean absolute error (*MAE*):

$$SDev = \sqrt{2}\cdot MAE = \frac{\sqrt{2}}{N_p}\sum_{N_p}|p_j - o_j|. \quad (3)$$

2.5.2. Flux errors

The observed flux o_i is the actual (true) value F plus a systematic error δ_o and a random error ϵ_o in the measurement (after Lasslop et al., 2008) for each measured half-hour i:

$$o_i = F_i + \delta_{o,i} + \epsilon_{o,i}\ . \quad (4)$$

The predicted (gap-filled) flux p_j can be stated accordingly where F is the true flux and δ_p and ϵ_p are the systematic and random error of the gap-filling technique for each predicted half-hour j:

$$p_j = F_j + \delta_{p,j} + \epsilon_{p,j} \quad (5)$$

The aggregated flux *FluxSum* is the sum of observed (measured) and predicted (gap-filled) fluxes:

$$FluxSum = \sum_{N_o} o_i + \sum_{N_p} p_j \quad (6)$$

with N_o denoting the number of observed half-hourly fluxes and N_p the number of predicted half-hourly fluxes. To calculate the total uncertainty *totUnc* of the aggregated fluxes:

$$totUnc = \delta + \epsilon\ , \quad (7)$$

the errors need to be propagated differently for systematic and random errors.

For systematic errors δ, the fluxes are biased above or below the true value (accuracy), hence these errors sum up over all gap-filled half-hours in an ordinary sum.

Since unknown biases in the measurements δ_o (e.g. through advection or instrument errors) cannot be retraced afterwards, only the systematic error of the predicted fluxes δ_p can be estimated for all gap-filled half-hours:

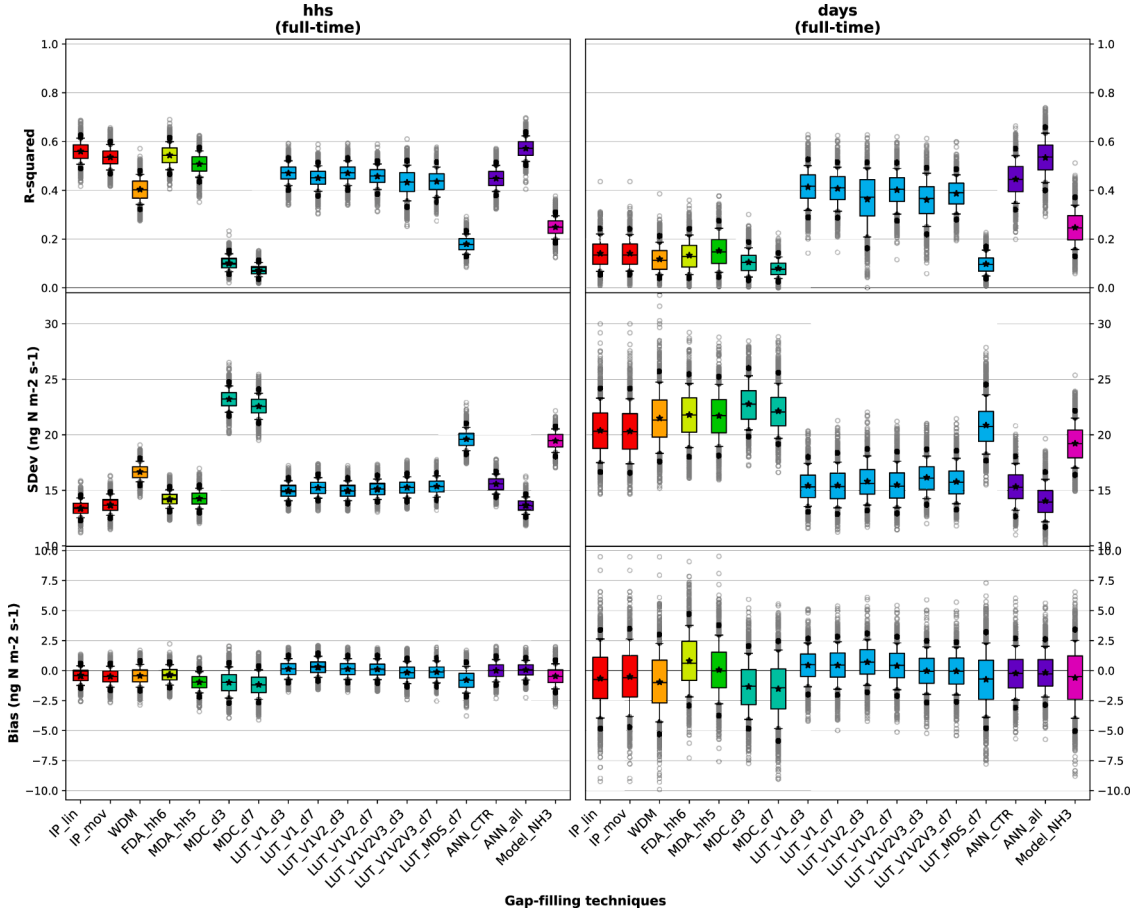


Fig. 3. Performance measures for the 'hhs' (*left*) and 'days' (*right*) scenarios of $\mathbf{NH_3}$ for all gap-filling techniques. The boxplot is composed of the median (solid line), mean (star symbol), lower and upper quartile bounds (box), 10th and 90th percentile (whiskers), and all outliers (dots) from the 999 bootstrapping samples.

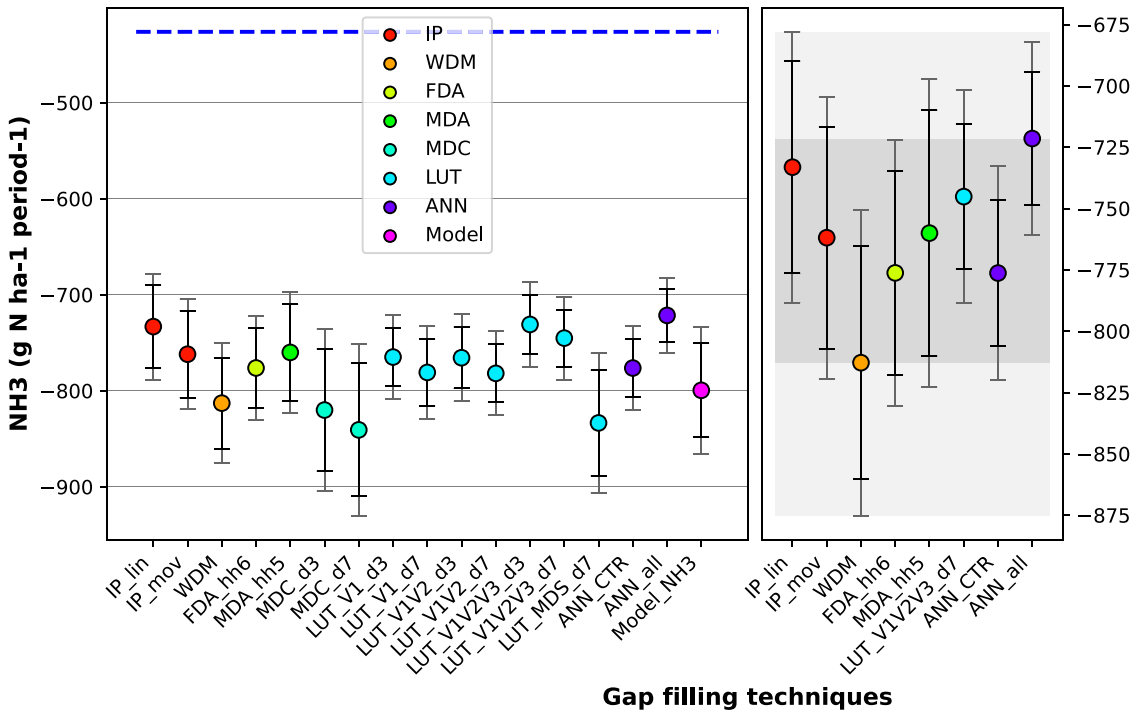


Fig. 4. Total sums of observed and predicted NH_3 fluxes for all gap-filling techniques (*left*) with the sum of the observed fluxes as a baseline (blue dotted line) and for the ensemble only (*right*). The sums are plotted with bias errors (black whiskers) and random uncertainties (gray whiskers) and with the range of the ensemble sums (dark gray box) and their uncertainties (light gray box).

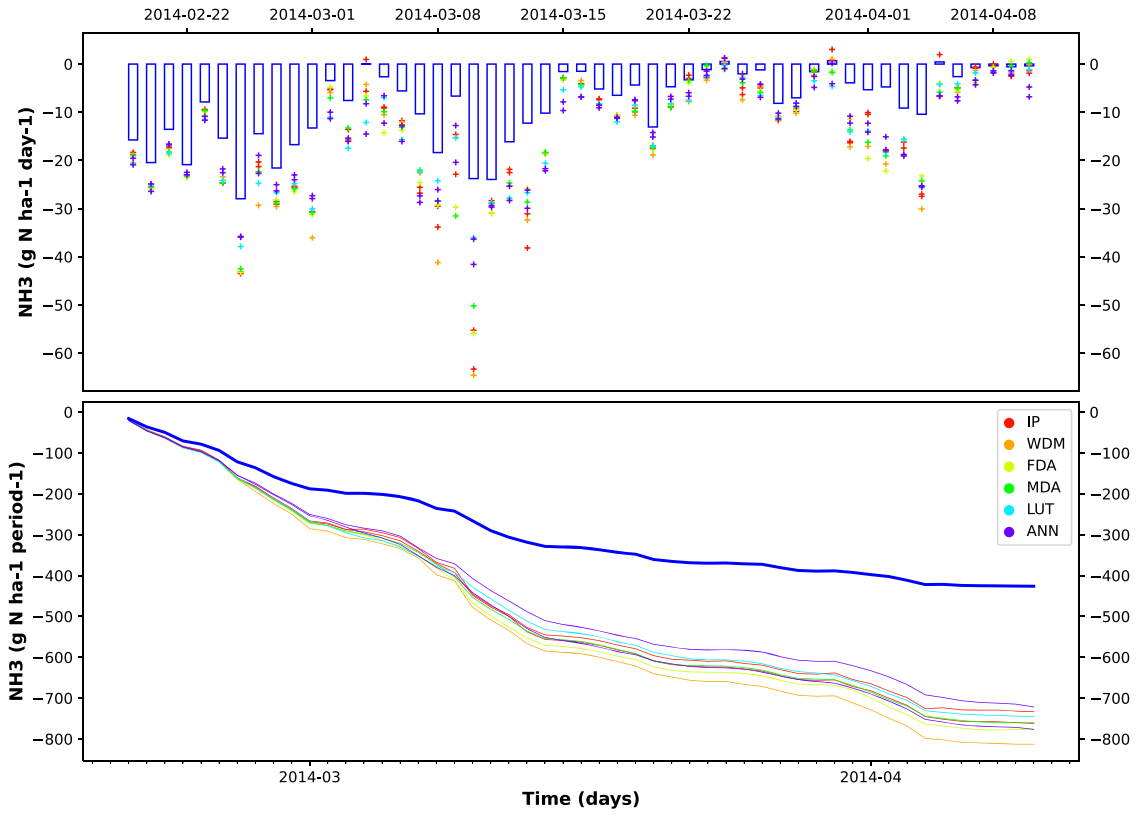


Fig. 5. Daily sums of observed (blue) and predicted NH_3 fluxes (*top*) and cumulated daily sums (*bottom*) for the gap-filling technique ensemble.

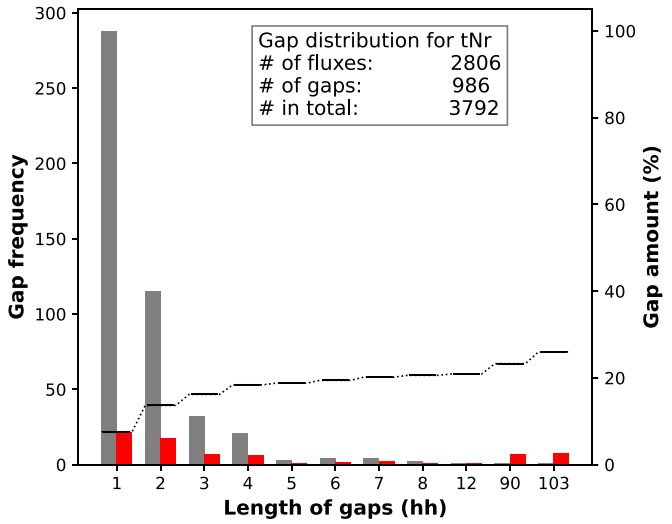


Fig. 6. Gap distribution for $\mathbf{tN_r}$ sorted by gap length. The histogram is drawn as in Fig. 2.

$$\delta \rightarrow \delta_p = \sum_{N_p} \delta_{p,j} \ . \tag{8}$$

For random errors ϵ, the fluxes are equally likely to be higher or lower than the true value (precision). The random errors add in quadrature, also called ordinary least squares. The total random error on the *FluxSum* is:

$$\epsilon = \sqrt{\sum_{N_o} \epsilon_{o,i}^2 + \sum_{N_p} \epsilon_{p,j}^2} \ . \tag{9}$$

As shown in Richardson et al. (2008) for CO_2, the statistical properties of the random uncertainty of the measurements ϵ_o can be inferred from the model residuals of the gap-filling techniques ϵ_p:

$$\epsilon_o \approx \epsilon_p \ . \tag{10}$$

To get an estimate of the uncertainties, the random uncertainty will be inferred using Eq. 10 also for the Non-CO_2 fluxes. This leads to a total random error of:

$$\epsilon \approx \sqrt{\sum_{N} \epsilon_{p,j}^2} \ . \tag{11}$$

The error propagation in Eq. 9 is assuming independent random errors. However, the random errors of the eddy covariance fluxes are probably auto-correlated as has been shown for CO_2 e.g. in Lasslop et al. (2008) and Menzer et al. (2013) using suitable statistical methods. When accounting for auto-correlation, the random error estimates were twice to three times higher than without auto-correlation (Menzer et al., 2013). Hence, calculating the errors from the model residuals without being able to account for measurement biases and for auto-correlations in errors, poses a lower limit on the estimated uncertainties.

2.5.3. *Bootstrapping*

Having secondary datasets with each half-hour as artificial gaps allows to explore any number of scenarios by bootstrapping sub-samples (Efron and Tibshirani, 1994). One bootstrap sample is one artificial gap scenario. Here, rather than having 10 fixed permutations as in

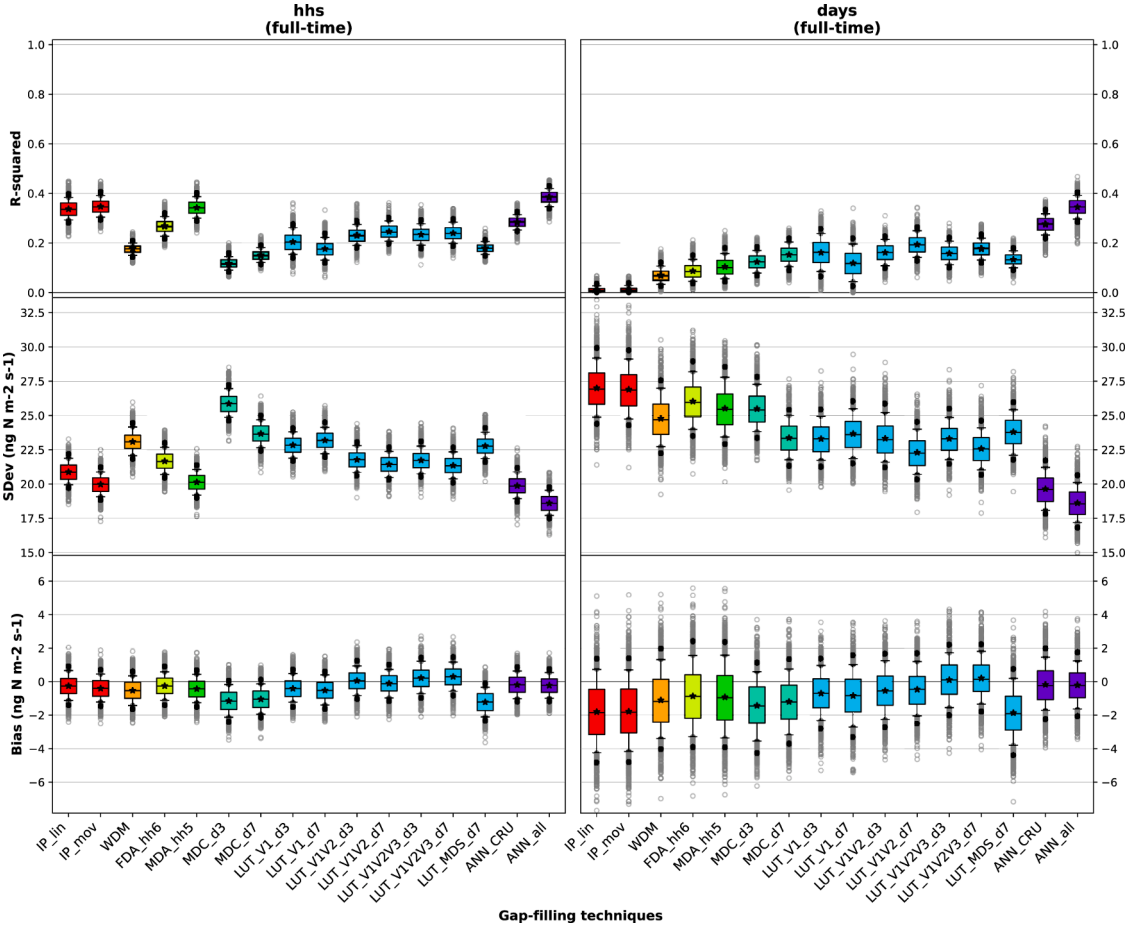


Fig. 7. Performance measures for the 'hhs' (*left*) and 'days' (*right*) scenarios of $\mathbf{tN_r}$ for all gap-filling techniques. The boxplot is drawn as in Fig. 3.

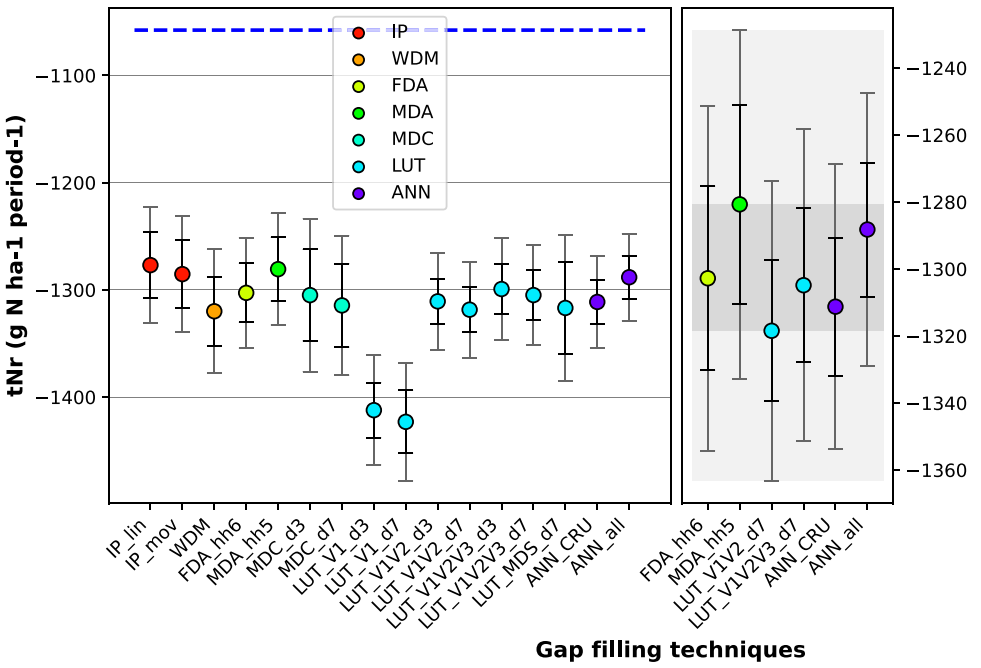


Fig. 8. Total sums of observed and predicted $\mathbf{tN_r}$ fluxes for all gap-filling techniques (*left*) with the sum of the observed fluxes as a baseline (blue dotted line) and for the ensemble only (*right*). For details on the plot see Fig. 4.

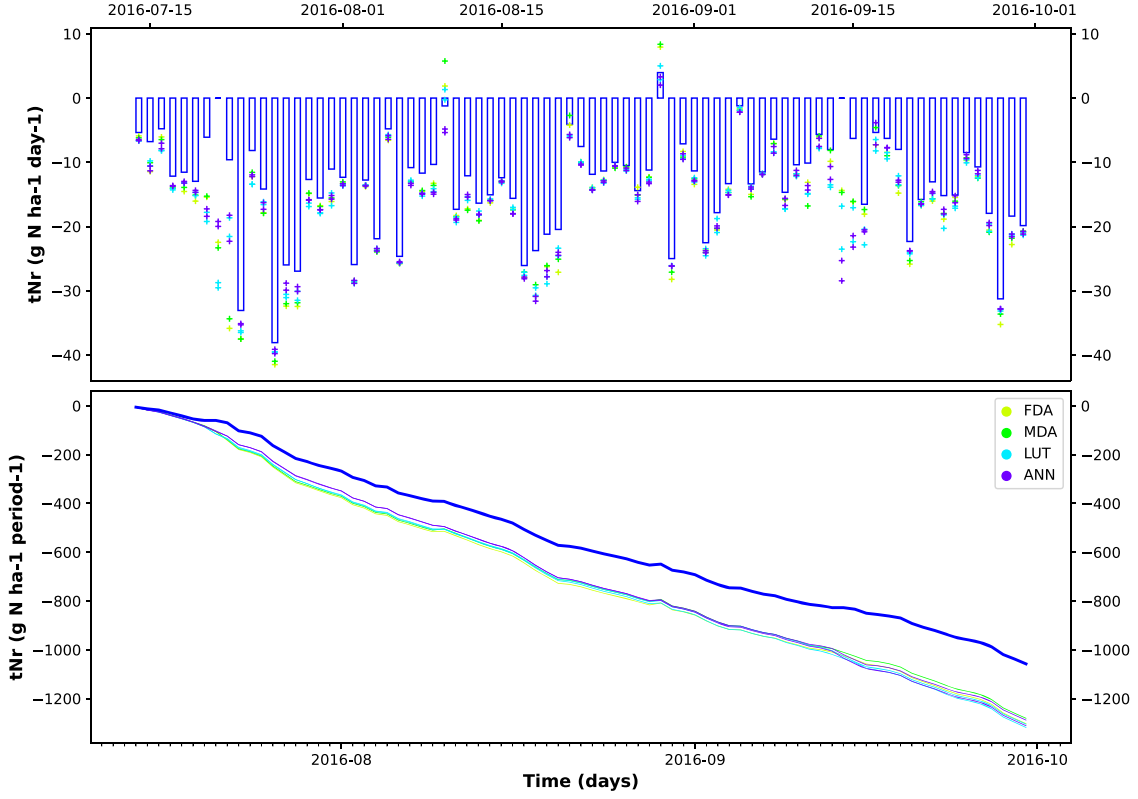


Fig. 9. Daily sums of observed (blue) and predicted $\mathbf{tN_r}$ fluxes (*top*) and cumulated daily sums (*bottom*) for the gap-filling technique ensemble.

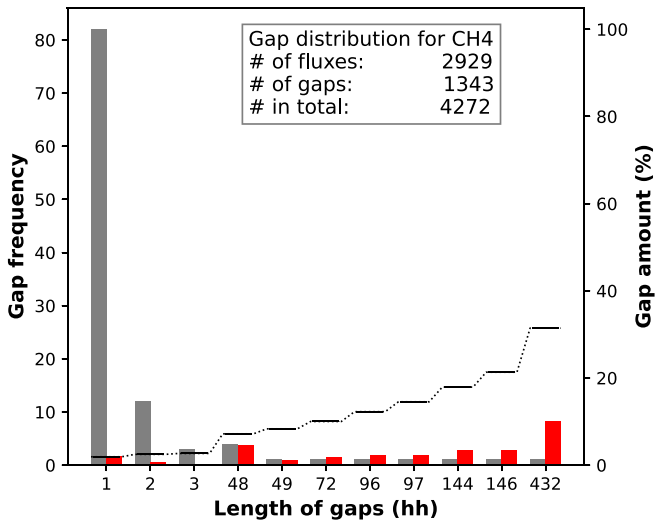


Fig. 10. Gap distribution for CH_4 sorted by gap length. The histogram is drawn as in Fig. 2.

Moffat et al. (2007), 999 artificial gap scenarios were randomly sampled with replacement. The size of the subsets affects the variance between bootstrapping samples; fifty percent was chosen as a rough typical percentage of gaps in a dataset.

When bootstrapping the 'hhs'- scenarios, randomly distributed half-hours were picked for the sub-samples. To include the effect of daily gaps in terms of placement and length, the half-hours of randomly distributed full days were picked for the 'days' scenarios. The bootstrapping was performed on three subsets of data: only 'daytime' data, only 'nighttime' data, and all data ('fulltime'). This resulted in two ('hhs', 'days') times three ('daytime', 'nighttime', 'fulltime') times 999 bootstrap samples. For each of these artificial gap scenarios, the three performance metrics described in Section 2.5.1 were calculated from the model residuals of the gap-filling technique.

As mentioned in Section 2.3.3, a few (longer) gaps could not be filled by some of the techniques. These data points were removed from the analysis in order to compare the same (sub)set of data across all techniques. Hence, only artificial gaps filled with the complete suite of gap-filling techniques were used for bootstrapping.

2.5.4. Uncertainties of the aggregated fluxes

To get estimates of the uncertainty for the aggregated fluxes, the total random uncertainty ϵ on the *FluxSum* (Eq. 11) was calculated from the mean of the standard deviation $\overline{SDev}$ of the model residuals obtained from bootstrapping artificial gap scenarios:[6]

$$\epsilon = \sqrt{N \cdot \overline{SDev}^2}. \tag{12}$$

Since the random error estimation requires an optimal model performance (where the deviation is mainly caused by the random error of the measurements), the $\overline{SDev}$ of the 'hhs' scenarios were used.

The systematic error δ of the gap-filled fluxes was calculated from 10% or 90% percentiles of the bootstrapping samples as in Moffat et al. (2007):

$$\delta = N_p \cdot \max(BE_{10}, BE_{90}). \tag{13}$$

To account for the influence of gap length, the bias error of the 'hhs' scenarios was taken for small gaps (≤12 half-hours) and the bias error from 'days' for all longer gaps.

[6] This equation is essentially the same as the square-root-of-time rule used in Vitale et al. (2019b) as an estimate of uncertainty.

2.5.5. Ensemble results of the sums

To calculate the sums of the measured and gap-filled fluxes over the whole period, the real gaps in the dataset were filled. These aggregated fluxes were calculated for all gap-filling techniques. The gap-filling techniques showing a medium to good performance in terms of high R^2, low *SDev*, and small *BE* centered around zero were used to calculate the ensemble results. (For *MDC* and *LUTs* with minor variants in window size, only the variant with ±7 days (*_d7*) was included in the ensemble of gap-filling techniques.)

Though the gap-filling techniques were pre-chosen to be based on different methods, the selection was still arbitrary. Besides, their workings are not independent and variants have been included. Therefore, general statistics such as calculating the median cannot be applied and the ensemble of results are stated as the range of the sums estimated by the ensemble gap-filling techniques t :

$$[\min(FluxSum_t),\ \max(FluxSum_t)]. \tag{14}$$

The difference between the lowest sum estimate and the highest sum estimate will be called *Delta* and used as a measure of the generic gap-filling error spanning multiple gap-filling techniques:

$$Delta = \max(FluxSum_t) - \min(FluxSum_t). \tag{15}$$

For the uncertainties, the overall lowest and highest limits of the techniques within the ensemble were used as the confidence intervals (CI):

$$\text{Lower limit CI} = \min(FluxSum_t - totUnc_t), \tag{16}$$

$$\text{Upper limit CI} = \max(FluxSum_t + totUnc_t). \tag{17}$$

Since the lower and upper limit of the sums and uncertainties are posed by different techniques, the lower and upper CIs are not symmetric.

3. Results and discussion

3.1. Trace gas specific gap-filling performance

The performance of the gap-filling techniques will first be evaluated separately for each of the four trace gases. The main criteria for a good gap-filling technique are a high R^2, low *SDev*, and small bias error centered around zero. Since the four different trace gas datasets have such different site and flux characteristics, the performance measures vary in absolute numbers, unit, and behavior and are not comparable between datasets. Therefore, the gap-filling performances have been rated individually in relative terms (best, medium, and low). The performances of the gap-filling techniques for the four trace gases can be found in Fig. 3, Fig. 7, Fig. 11, and Fig. 15. Details on the gap-filling of the artificial 'hhs'- and 'days'-scenarios such as time series of the observed and predicted fluxes, error measures when bootstrapping with the dataset split into daytime and nighttime data, or scatterplots for each gap-filling technique can be found in the supplements.

For NH_3 (Fig. 3), *LUT_V1V2V3, ANN_CTR,* and *ANN_all* have the best performance. These three techniques explain up to 60% of the variability in the fluxes for the 'hhs'- and 'days'-scenarios, a mean standard deviation around 15 ng N m^{-2} s^{-1}, and a bias error with 10/90-percentiles inside ±2.5 ng N m^{-2} s^{-1} centered around zero. *ANN_all* performs better than *ANN_CTR* hinting at (hidden) relationships with other drivers.

Medium performance is shown by the following techniques: *IP_lin, IP_mov, WDM, FDA*, and *MDA* work well for the 'hhs'-scenarios but have a lower R^2 and higher *SDev* for the 'days'-scenarios.

Low performances with lowest R^2, highest *Sdev*, and clear bias offsets are exhibited by *MDC* and *LUT_MDS* for the 'hhs'- and 'days'-scenarios when split into daytime and nighttime during bootstrapping (see Fig. S1.3 and Fig. S1.7). *LUT_V1* and *LUT_V1V2* also have quite large offsets in the bias errors when the dataset has been split.

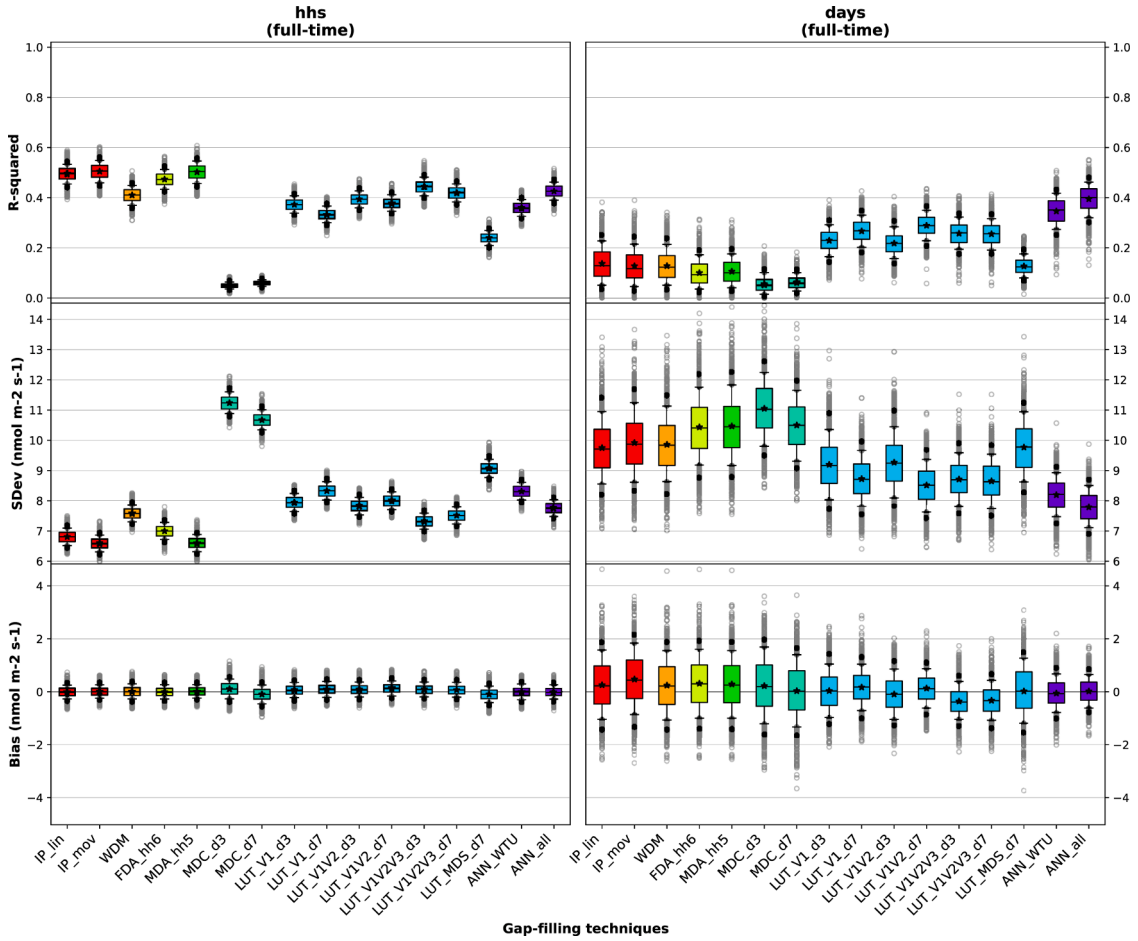


Fig. 11. Performance measures for the 'hhs' (*left*) and 'days' (*right*) scenarios of CH_4 for all gap-filling techniques. The boxplot is drawn as in Fig. 3.

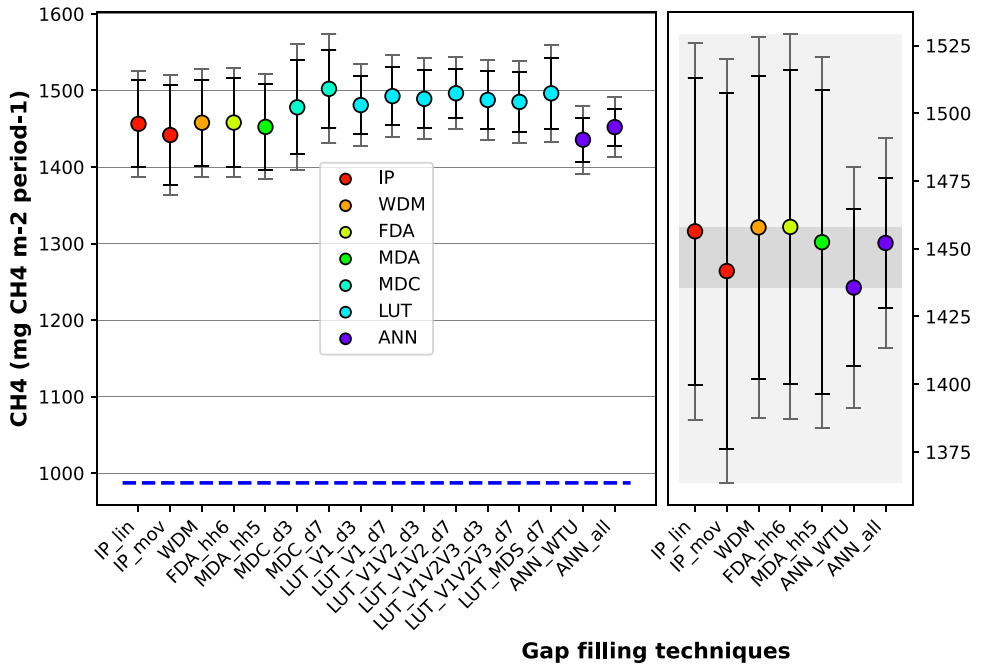


Fig. 12. Total sums of observed and predicted CH_4 fluxes for all gap-filling techniques (*left*) with the sum of the observed fluxes as a baseline (blue dotted line) and for the ensemble only (*right*). For details on the plot see Fig. 4.

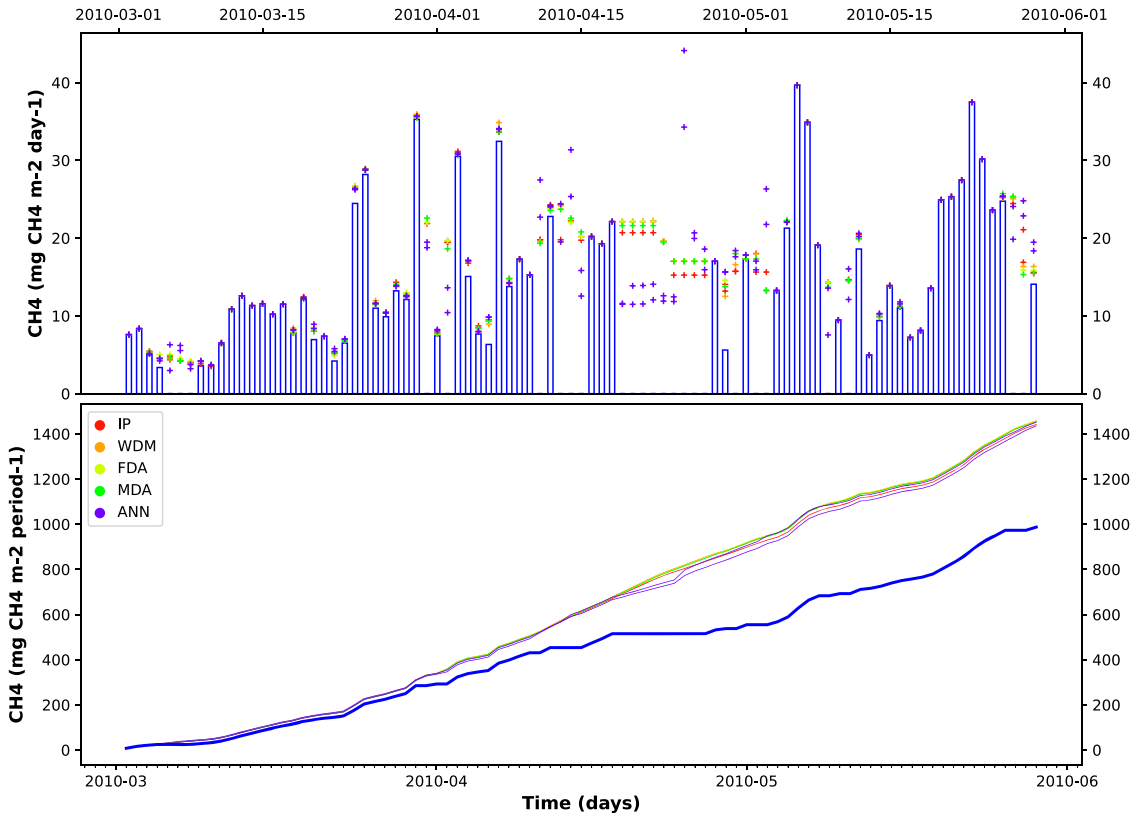


Fig. 13. Daily sums of observed (blue) and predicted $\mathbf{CH_4}$ fluxes (*top*) and cumulated daily sums (*bottom*) for the gap-filling technique ensemble.

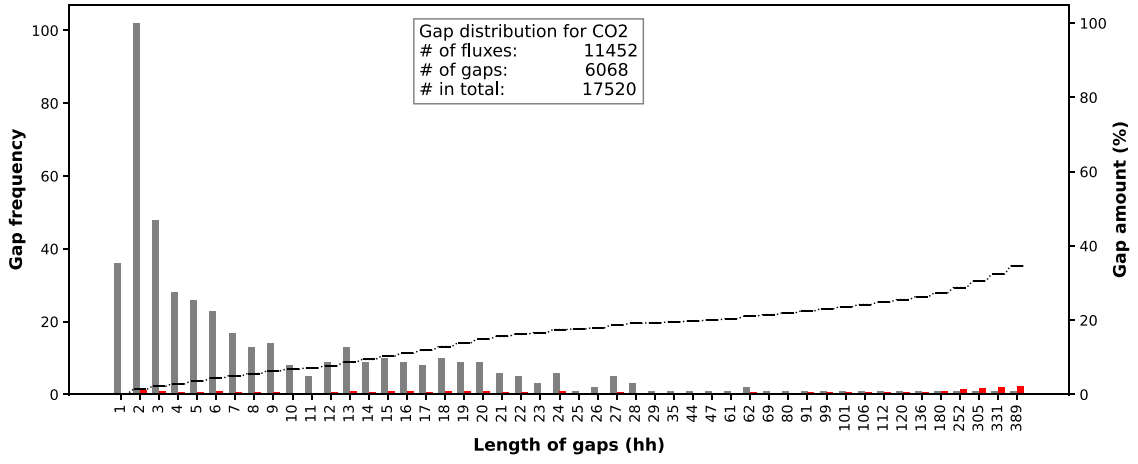


Fig. 14. Gap distribution for $\mathbf{CO_2}$ sorted by gap length. The histogram is drawn as in Fig. 2.

For tN_r (Fig. 7), the differences between techniques are less pronounced and even the best technique, *ANN_all*, explains only up to 40% of the variability. Overall, the two ANNs and the two *LUT_V1V2V3* show the best performance for the 'hhs'- and 'days'-scenarios with a mean standard deviation between 17 and 25 ng N m^{-2} s^{-1} and bias errors with 10/90-percentiles inside ±2 ng N m^{-2} s^{-1} centered around zero. *ANN_all* performs notably better than *ANN_CRU* hinting again at other (hidden) relationships.

Medium performances have *LUT_V1V2, FDA*, and *MDA* with a small but clear negative offset in the bias of the 'days' scenario. The R^2 performances are quite low for both, the 'hhs' and 'days' scenarios.

Low performance is observed for the *IP_lin, IP_mov, WDM, MDC* and *LUT_MDS* with clear negative bias offsets. *LUT_V1* with the concentration of tN_r as the only independent variable has almost no bias offsets if all the data is bootstrapped, but when looking at daytime and nighttime separately, there is a strong bias in the 'hhs'- and 'days'-scenarios (Fig. S2.3 and Fig. S2.7). The offsets disappear if radiation is included as an additional LUT variable which indicates that tN_r fluxes display diurnal patterns. Though concentration measurements can be used for gap-filling, radiation also needs to be included as a controlling variable (see also Zöll et al., 2019).

For CH_4 (Fig. 11), the best overall performance is achieved by *ANN_WTU* and *ANN_all* with similarly good results for the 'hhs' and 'days'-scenarios. The two techniques could explain 30% to 45% of the variability in the fluxes, with a mean standard deviation between 7.5 and 9 nmol m^{-2} s^{-1} and a bias error with 10/90-percentiles inside ±1

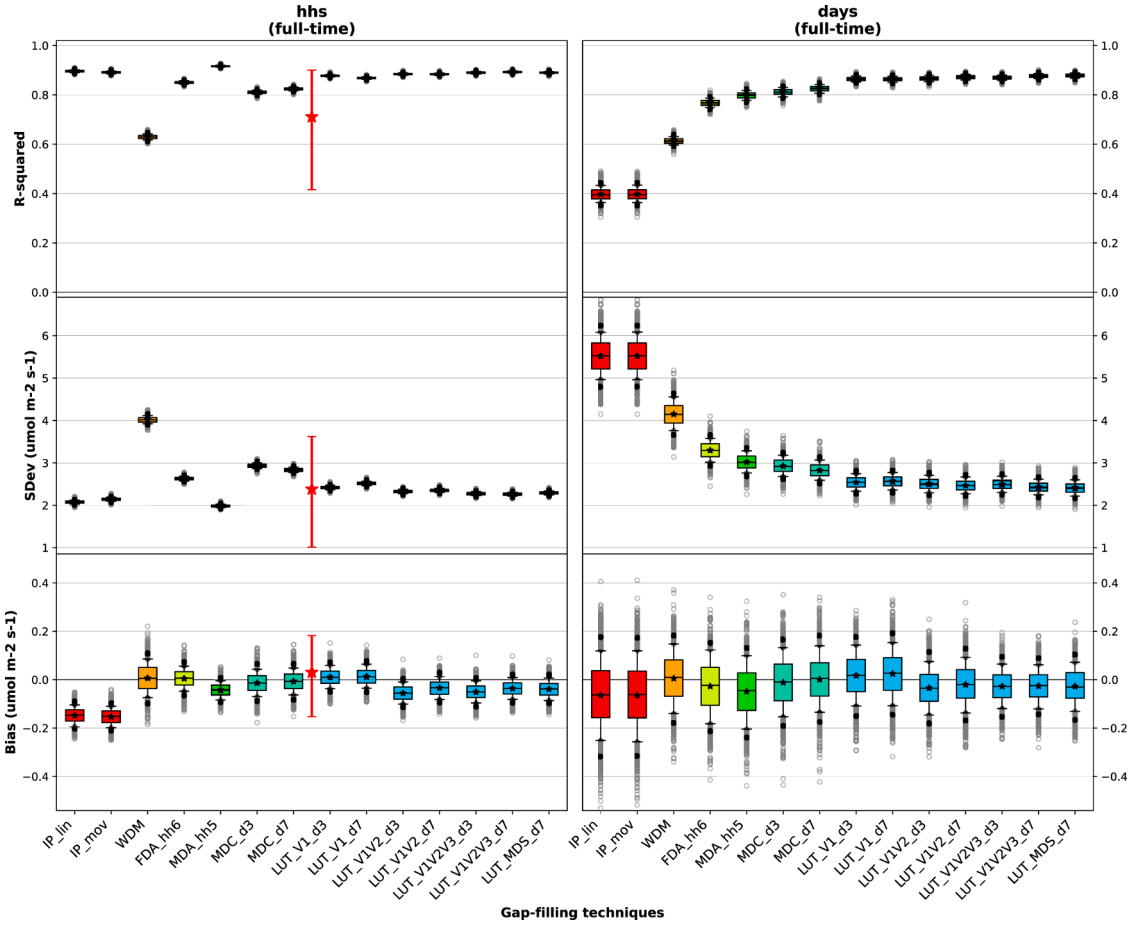


Fig. 15. Performance measures for the 'hhs' (*left*) and 'days' (*right*) scenarios of $\mathbf{CO_2}$ for all gap-filling techniques. The boxplot is drawn as in Fig. 3. For comparison, the results from (Moffat et al., 2007) of their very short artificial gap scenarios of single half-hours have been added to the 'hhs' scenarios (*left*): the mean (red star) and 10th- and 90th-percentiles (red whiskers) over 10 scenarios and 18 gap-filling techniques.

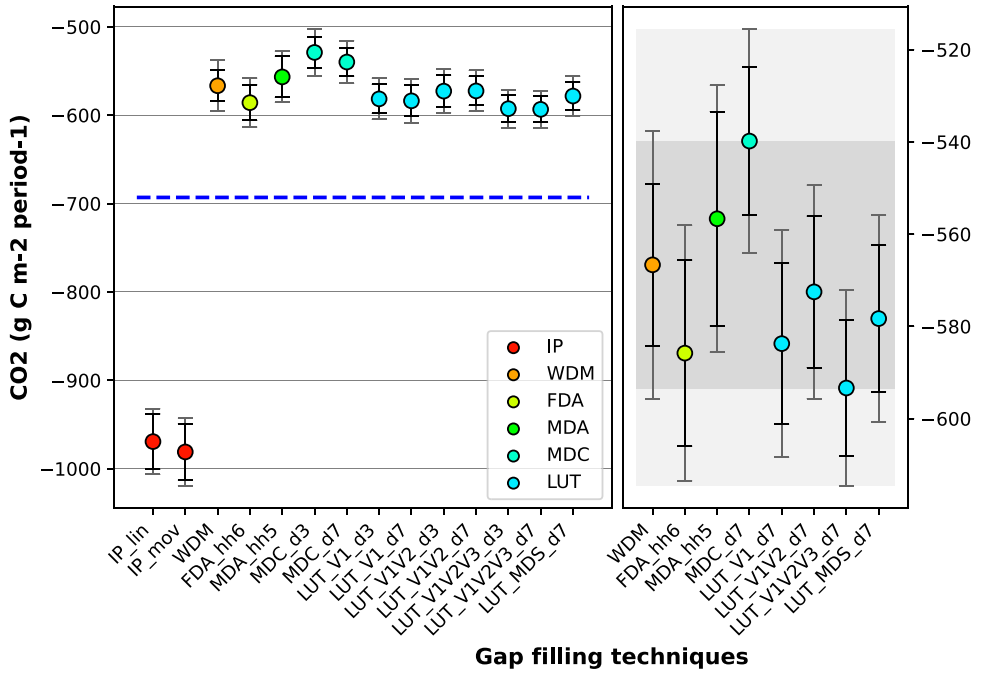


Fig. 16. Total sums of observed and predicted $\mathbf{CO_2}$ fluxes for all gap-filling techniques (*left*) with the sum of the observed fluxes as a baseline (blue dotted line) and for the ensemble only (*right*). For details on the plot see Fig. 4.

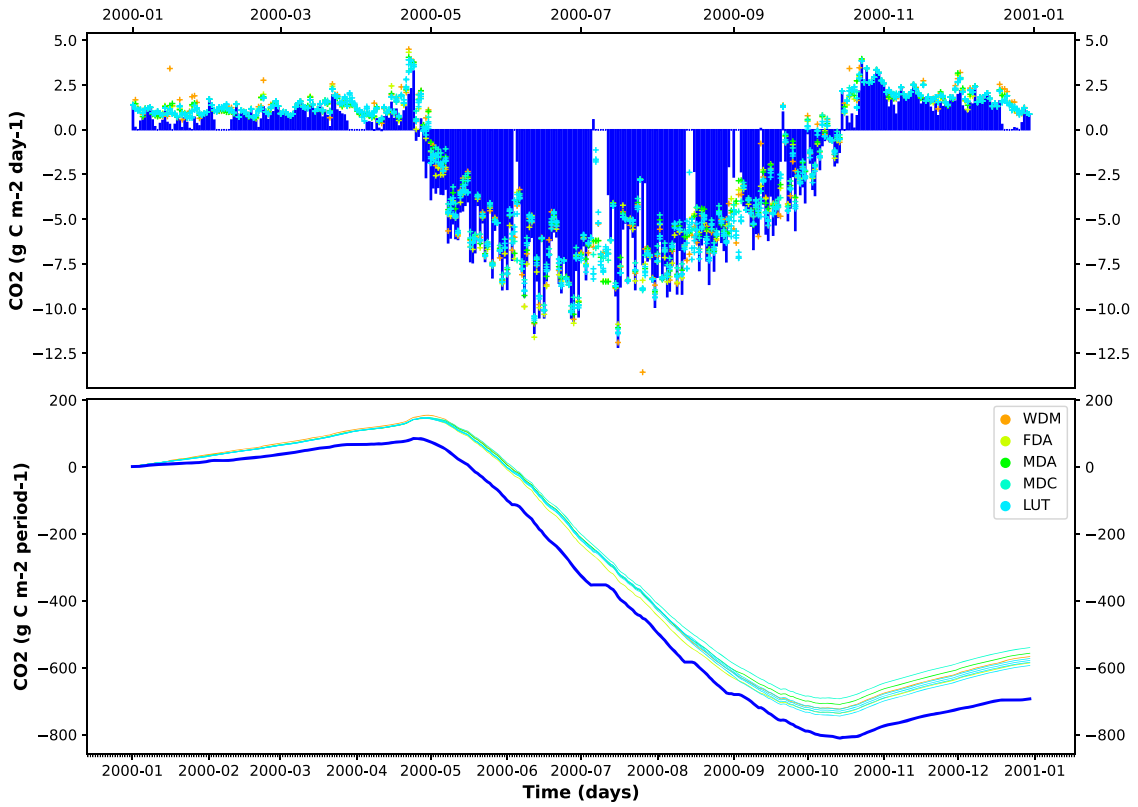


Fig. 17. Daily sums of observed (blue) and predicted NH_3 fluxes (*top*) and cumulated daily sums (*bottom*) for the gap-filling technique ensemble.

nmol m^{-2} s^{-1}.

Medium performances have *IP_lin, IP_mov, WDM, FDA*, and *MDA*. Though in the 'hhs'-scenarios all five techniques even have slightly better performance metrics than the two ANNs (highest R^2, lowest *Sdev* and small bias errors centered around zero), their performance is below average in the 'days'-scenarios. The good performance in the 'hhs'-scenarios of these interpolation techniques indicates that the fluxes are gradually changing with little fast responses to driving factors.

Low performances are found for *MDC* and *LUT_MDS* with the lowest R^2 and highest *Sdev*. The fact that *LUT_MDS* has small bias errors but such a low R^2 indicates that the diurnal pattern in the fluxes is not due to radiation but other drivers (such as *ustar* which reduces the bias offsets if taken as the third variable in *LUT_V1V2V3*). When the dataset is split into daytime and nighttime during bootstrapping (Fig. S3.3 and Fig. S3.7), *LUT_V1, LUT_V1V2, LUT_V1V2V3,* and also *ANN_WTU* exhibit large offsets in the bias errors.

For CO_2 (Fig. 15), over 80% of the variability in the fluxes can be explained by the gap-filling techniques. Most techniques have such a high R^2 for the 'hhs'- and 'days'-scenarios with a mean standard deviation below 4 umol m^{-2} s^{-1}. The bias errors have 10/90-percenticles inside ±0.2 umol m^{-2} s^{-1} but all variants of LUT exhibit clear offsets if split into daytime and nighttime for bootstrapping (Fig. S4.3 and Fig. S4.7). The R^2 and *SDev* performances of *WDM* are quite low but it is the only technique with bias offsets centered around zero for the daytime and nighttime subsets.

The two techniques *IP_lin* and *IP_mov* have a low performance with clear bias offsets which are even larger if split into daytime and nighttime for the 'days'-scenarios (Fig. S4.7).

3.2. General gap-filling performance

Though the performance of each gap-filling technique differs for the four different trace gas datasets (Fig. 3, Fig. 7, Fig. 11, Fig. 15), some of the findings can be attributed to the workings of the technique.

The two simple linear interpolation techniques, *IP_lin* and *IP_mov*, can be recommended for filling small gaps (i.e. several half-hours) in

Table 1
Overview of gap-filling specific properties of the datasets and variables used for the look-up tables (LUTs, Section 2.3.3) and for the artificial neural networks (ANNs, Section 2.3.4) using the same three inputs as in the LUTs. Naming of *ANN_XYZ* based on first letters of the three variables used.

Properties	NH_3	tN_r	CH_4	CO_2
Days of (half-hourly) data	51	79	89	365
Percentage of gaps	54.5	26.0	31.4	34.6
Variable 1	C_NH3	C_Nr	Wind_Direc	Rg
LUT bin width	*±5.0 ppb*	*±0.6 ppb*	*±22.5°*	*±50 W m^{-2}*
Variable 2	Rg	Rg	Ts_20	Ta
LUT bin width	*±50 W m^{-2}*	*±50 W m^{-2}*	*±1.25°C*	*±2.5°C*
Variable 3	Ts_20	Ustar	Ustar	VPD
LUT bin width	*±1.25°C*	*±0.1 m s^{-1}*	*±0.1 m s^{-1}*	*±2.0 hPa*
ANN naming	*ANN_CRT*	*ANN_CRU*	*ANN_WTU*	—

(C_NH3 – concentration of NH_3, C_Nr – concentration of total reactive nitrogen, Rg – global radiation, Ta – air temperature, Ts_20 – soil temperature in 20 cm depth, Ustar – friction velocity, VPD – vapour pressure deficit).

Table 2
Overview of gap-filling techniques.

Gap-filling technique	Abbreviation	Short description
Simple interpolations (Section 2.3.1):		
Linear interpolation	*IP_lin*	Linear interpolation between previous and next existing data point
Moving average	*IP_mov*	Moving average of five half-hours at a time
Diurnal interpolations (Section 2.3.2):		
Weighted daylight mean	*WDM*	Mean of the daylight and mean of nighttime half-hours
Fixed diurnal average	*FDA*	Average in steps of fixed half-hours (*FDA_6hh,* 0:00–03:00, …, 21:00–24:00)
Moving diurnal average	*MDA*	Moving average of five half-hours at a time (*MDA_5hh*), extending to adjacent days
Mean diurnal course	*MDC*	Average of single half-hours at the same time of day for consecutive days, increasing in steps of ±3 days (*MDC_d3*) or ±7 days (*MDC_d7*)
Look-up tables (Section 2.3.3):		
Look-up table	*LUT*	Binning depending on one, two, or three variables (*LUT_V1, LUT_V1V2, LUT_V1V2V3,* see also Table 1) within a certain time window of consecutive days, increasing in steps ±3 days (*_d3*) or ±7 days (*_d7*)
	LUT_MDS	Binning depending on the same variables and ranges as the main look-up table of *MDS*
Artificial neural networks (Section 2.3.4):		
Artificial neural network	*ANN*	Artifical neural networks with three variables as inputs (*ANN_XYZ*, see also Table 1) and with all available variables (*ANN_all*)
Inferential model (Section 2.3.5):		
Ammonia flux model	*Model_NH3*	Inferential model for the biosphere-atmosphere exchange of ammonia

datasets with fluxes with little or no diurnal cycle and non-sporadic fluxes. However, for fluxes with a strong diurnal cycle such as CO_2, the two techniques are not suitable, exhibiting strong offsets in the bias errors. For the three other trace gases, these two techniques are among the best in the 'hhs'-scenarios (high R^2, low *SDev*, and small, centered bias errors). For NH_3 and CH_4, they even show a medium performance for the 'days'-scenarios (though with a strong drop in performance compared to 'hhs').

The *WDM* technique has been included as a new gap-filling technique. The method is too simple to capture the variability of the fluxes. However, the bias error is as small or even a bit smaller than for the other diurnal interpolation methods for all four trace gas dataset, for the 'hhs'- as well as the 'days'-scenarios, and also if separated into daytime and nighttime for bootstrapping. Since a low and centered bias error is crucial for a reliable aggregation of the fluxes, the weighted diurnal mean might be a useful technique for quick estimates of daily sums and can even be recommended for annual sums especially for trace gases with a diurnal cycle.

The diurnal interpolation techniques differ in their performances. *FDA* and *MDA*, are among the best techniques for the 'hhs'-scenarios and medium for the 'days'-scenarios. In contrast, the performance of *MDC* is lowest for all three Non-CO_2 trace gases (low R^2, high *SDev*, clear bias error offsets). This technique cannot be recommended as a gap-filling technique for noisy fluxes with little diurnal cycle. The two variants of *MDC* with a window size of ±3 and ±7 days show only little differences.

The fact that the performance of the simple interpolations, *IP_lin* and *IP_mov*, is better than the diurnal interpolation with *MDC* means that the correlation of the missing data (gap) to the preceding and succeeding half-hours (basic principle of *IP_lin, IP_mov*) is higher than to the half-hour at the same time of day from the preceding and succeeding day (basic principle of *MDC*). By combining these two principles, the *MDA* technique matches or outperforms the other interpolation methods for all datasets. The performance of *FDA* is comparable to *MDA* and its good performance indicates that *FDA* may be deployed to reduce the noise and fill the gaps in the data at the same time.

The LUTs work well (high R^2, low *SDev*, small Bias) for the 'hhs'- and 'days'-scenarios on all four datasets when the independent variables are pre-selected to have high correlations (see Section 2.3.3). The differences between *LUT_V1V2* and *LUT_V1V2V3* are small, though using only one variable, *LUT_V1*, may yield unstable results. Again, little difference in performances is observed between the variants ±3 and ±7 days.

As the *MDS* algorithm has been developed for CO_2, *LUT_MDS* and *LUT_V1V2V3* with the same set of standard variables[7] show the best LUT performances for CO_2. However, for the other three trace gases, *LUT_MDS* exhibits the lowest R^2 of all LUT, the highest *SDev* and the largest bias error offset for tN_r. Equivalent findings are reported using the *MDS* algorithm with different predictor subsets for CH_4 (Irvin et al., 2021; Kim et al., 2019). These results should be taken into account when using the online REddyProc with default variable settings[2] for Non-CO_2 trace gases.

The artificial neural networks with all drivers (*ANN_all*) outperform the other gap-filling techniques for the 'hhs'- as well as 'days'-scenarios. They were used for the Non-CO_2 trace gases herein and have already been shown to perform best for CO_2 (Moffat et al., 2007). The ANNs with a subset of drivers perform almost as well. The subsets were based on the same three drivers as *LUT_V1V2V3*. The two types of techniques (LUTs and ANNs) show similarly high performance for the 'hhs'-scenarios of CH_4 and NH_3 but the ANNs are better for all the 'days'-scenarios highlighting their strength for longer gap sizes. The overall good performance of ANNs is very promising since the exact drivers of Non-CO_2 trace gases can often not be generalized due to differences in atmospheric composition, gas reactivity, and ecosystem dependent characteristics. For CH_4, a variety of machine learning algorithms including ANNs have already been successfully deployed (e. g. Irvin et al., 2021; Kim et al., 2019).

The NH_3 dry deposition inferential model exhibits the largest bias error of all techniques and relatively low performance in the other metrics (comparable to the interpolation-based techniques). To a certain degree, a large bias error can be expected since the model was applied

[7] The set of variables was pre-determined using ANNs. Only for CO_2, this yielded the same of the standard *MDS* variables (radiation, temperature, and VPD) with a small difference between LUT_V1V2V3_d7 and LUT_MDS_d7 in the setting of the VPD range of ±2.0 versus ±5.0.

Table 3
Ensemble results of the aggregated fluxes and confidence intervals (CI) for all four trace gases. (Detailed sums and uncertainty estimates can be found in the according supplements, **Table S1.1** for NH_3, **Table S2.1** for tN_r, **Table S3.1** for CH_4, and **Table S4.1** for CO_2).

Trace gas **Ensemble results**	**NH_3** (g N ha^1 period^{-1})	**tN_r** (g N ha^{-1} period^{-1})	**CH_4** (mg CH_4 m^{-2} period^{-1})	**CO_2** (g C m^{-2} period^{-1})
Upper limit CI	-678	-1229	+1529	-516
Upper uncertainty	*(+44)*	*(+52)*	*(+71)*	*(+24)*
Upper limit sum	**-722**	**-1281**	**+1458**	**-540**
Delta	*(91)*	*(37)*	*(22)*	*(53)*
Lower limit sum	**-813**	**-1318**	**+1436**	**-593**
Lower uncertainty	*(-62)*	*(-45)*	*(-72)*	*(-22)*
Lower limit CI	-875	-1363	+1364	-615
Total CI	*197*	*134*	*165*	*99*

without calibration using land-use specific parameters originally derived to be valid for a broad range of semi-natural ecosystems (Massad et al., 2010). Site-specific calibrations of individual model parameters may increase the value of this model as a gap-filling tool and help reduce the bias error (Schrader et al., 2016; Schrader et al., 2020).

The CO_2 flux dataset has been included here for comparison with Moffat et al. (2007). Therein, eighteen techniques were tested on different artificial gap scenarios including ten 'very short'-scenarios of single half-hours. The results for this specific scenario and dataset were added to the 'hhs'-scenarios of Fig. 15. All techniques but the two linear interpolation methods not suitable for CO_2 are the same or similar techniques as in Moffat et al. (2007). The performances of these techniques (R^2, *SDev*, and bias error) are inside the ranges of Moffat et al. (2007) demonstrating that new artificial gap-filling scheme yields comparable results.

3.3. Aggregated flux ensembles

For each trace gas dataset, the real gaps were also filled with all techniques, the fluxes aggregated over the full time period and the sums calculated for the ensemble of medium and best gap-filling techniques (Table 3, with details on the calculations in Section 2.5.5).

The NH_3 fluxes summed up over the period of 51 days have a *Delta* of 91 g N ha^{-1} period^{-1} for the gap-filling ensemble and a total confidence interval of 197 g N ha^{-1} period^{-1} (Fig. 4). The variability of the sums of the medium and best techniques is thus about 10% total NH_3-N deposition. The prediction of the NH_3 dry deposition inferential model is inside this range. Looking only at the best techniques would reduce the *Delta* to 55 g N ha^{-1} period^{-1} but comprise only two types of gap-filling techniques with LUT and ANN related in their working principles.

For tN_r, the fluxes over the period of 79 days yield to a *Delta* of 37 g N ha^{-1} period^{-1} and a total confidence interval of 134 g N ha^{-1} period^{-1} for the medium and best gap-filling techniques (Fig. 8). Due to the strong bias error offsets of *LUT_V1* for daytime and nighttime, the estimated sum is outside the confidence interval of the ensemble.

The CH_4 fluxes aggregated over 89 days with the gap-filling ensemble result in a *Delta* of 22 mg CH_4 m^{-2} period^{-1} and a total confidence interval of 165 mg CH_4 m^{-2} period^{-1} (Fig. 12). The estimated sums are all similar despite differences in gap-filling performances including techniques with clear offsets in the bias errors and despite the fact that a high percentage of the gaps came from full day gaps of up to nine days (Fig. 10 and Fig. 13). This may again be attributed to the more gradually changing, mostly positive CH_4 fluxes at this site. The *Delta* across all techniques is 66 mg CH_4 m^{-2} period^{-1}. Its size is within the range of the standard deviation of differences in annual sum estimates for ten gap-filling techniques reported by Kim et al. (2019, Table 4 therein) of 42 to 262 mg CH_4 m^{-2} period^{-1} depending on the site and year with periods of 6 to 12 months.

For CO_2, fluxes from a full year of measurements were filled (Fig. 16). The *Delta* of the gap-filling ensemble was 53 g C m^{-2} period^{-1} and the total confidence interval 99 g C m^{-2} period^{-1}. The bad performance of the simple interpolation techniques shows how crucial it is to account for the diurnal cycle of CO_2. In contrast to the other three trace gases, the differences between techniques are larger than their error bounds which could partly be an effect of the longer dataset. Moffat et al. (2007) reported a more generic gap-filling error of 0.25 g C m^{-2} per gap-filled day derived from the spread of the bias error of the well-performing gap-filling techniques investigated therein. Applying this generic gap-filling error to the CO_2 dataset translates to potential differences in annual sum estimates between gap-filling techniques of ±31.6 g C m^{-2}, hence a range of 63.2 g C m^{-2}. This range has a similar size as the range of the annual sums estimates of the ensemble gap-filling techniques (i.e. the *Delta*) found herein. The *Delta* (Eq. 15) is an alternative method of expressing such a generic gap filling error spanning multiple techniques which can be calculated dataset specific for any kind of trace gas.

The only difference between *LUT_MDS* and *LUT_V1V2V3_d7* is the setting for the bin width of VPD (5 hPa and 2 hPa, respectively) and led to a difference in the annual sum of 15 g C m^{-2} (Figure 16) and Table S4.1). This demonstrates how even small changes in the setup of the *same* gap-filling technique can lead to significant differences in the annual sum and underlines the need for a more robust general gap-filling approach.

By using multiple gap-filling techniques, the aggregated fluxes will be less dependent on the choice of the techniques (and their settings) and the estimated uncertainties will be more robust. The ensemble sums reported herein are stated as ranges (Table 3) since the suite of gap-filling techniques used was arbitrary and limited in methods (see Section 2.5.5). However, reporting ranges may even be more representative than absolute values for inter-annual and inter-site comparisons since the uncertainties induced by gap-filling are highly dataset specific (see also Section 1).

Improving the gap-filling of eddy covariance fluxes through using ensembles has recently also been suggested in Mahabbati et al. (2021). As in other fields of Earth Science, this research can profit from the knowledge already obtained in climate modelling with similar benefits and challenges as discussed for crop modelling in Wallach et al. (2016).

To minimize the effect of the choice of gap-filling techniques, the suite of techniques should encompass a wide range of different gap-filling approaches. The gap-filling framework developed herein had a focus on campaign data of a few weeks to months. However, our framework is just one potential implementation of the proposed

methodology (Section 2.1). In a way, the chosen gap-filling techniques and example datasets are only a case study on how it can be realized. For long-term measurements and multi-site studies, long gaps play a more crucial role. Therefore, the gap-filling framework should include an artificial gap scheme with longer gap lengths (e.g. weekly) and more machine learning techniques as these have shown to be advantageous for filling long gaps (e.g. Kim et al., 2019; Zhu et al., 2022). In cases where absolute numbers are needed such as in data assimilations or meta analyzes, further research is required to test the applicability of different ensemble metrics (mean, median, center, weighted mean etc.) while deploying a wide variety of gap-filling techniques.

4. Conclusions and outlook

The results show that the implemented gap-filling framework was successfully applied to four different trace gases. The new artificial gap scheme has the major advantage that as much information of the dataset as possible is preserved during the artificial gap-filling procedure while obtaining complete secondary datasets. The artificial gap scenarios are then sub-sampled by bootstrapping. The statistical metrics calculated from the model residuals of the gap-filling techniques are used to evaluate the performances of the gap-filling techniques as well as to quantify the random and systematic errors.

The techniques with acceptable performance were then used to gap-fill the real gaps and calculate an ensemble of aggregated fluxes. The development of such flexible gap-filling tools based on multiple gap-filling techniques is essential for Non-CO_2 trace gases to identify suitable techniques. The suite of gap-filling techniques may be extended to capture an even wider range of dataset characteristics such as the sporadic nature of N_2O fluxes. For non-aggregated fluxes in cases where complete gap-filled times series are required directly on the half-hourly basis, further research is needed assessing the implications of gap-filling with single versus multiple techniques.

The new methodology describes a universal gap-filling framework based on multiple gap-filling technique and can be adopted for any kind of eddy covariance datasets. It encompasses the key elements needed for standardizing gap-filling procedures: a) an evaluation of the gap-filling techniques, b) an assessment of their uncertainties, c) a quantification of the more generic gap-filling error spanning multiple gap-filling techniques accounting for the specific dataset characteristics, and d) ensemble estimates of the aggregated fluxes.

We recommend that the standardized processing of gap-filling eddy covariance datasets should be based on ensemble results from multiple techniques as this will increase the significance of inter-site comparisons. Aggregating fluxes such as daily or annual sums from an ensemble of gap-filling techniques will be an important step towards more transparency and rigor. By implementing the key elements of the proposed methodology in the standardized gap-filling pipelines of research infrastructures such as ICOS, seasonal to multi-year budget estimates will become less biased towards a single gap-filling technique and uncertainty estimates will become more robust and defendable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The original code and one example datetset are provided at the Carbon Portal DOI minting service (https://doi.org/10.18160/R6S5-47J1).

Acknowledgements

Funding for this study was provided by the German Federal Ministry of Education and Research (BMBF) within the framework of the Junior Research Group NITROSPHERE under support code FKZ 01LN1308A and by the German Environment Agency (UBA) through the FOREST-FLUX project under support code FKZ 3715512110. Further funding was received through the project RINGO ("Readiness of ICOS for Necessities of integrated Global Observations") from the EU's Horizon 2020 research and innovation programme under grant agreement No 730944.

We thank Jeremy Rüffer and Jean-Pierre Delorme for their excellent technical support during the Bourtanger Moor and Bavarian Forest field campaigns, Undine Zöll and Pascal Wintjen for their support and scientific exchange regarding field campaign data, Ute Karstens and Claudio D'Onofrio for their support providing tool and code via the ICOS Carbon Portal. We further thank the two anonymous referees for their valuable comments that helped improve clarity and readability of the paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.agrformet.2022.109114.

Appendix A.1: Multiple Gap-Filling Tool

The multiple gap-filling tool (MGF-Tool) can be used for gap-filling fluxes in any kind of eddy covariance dataset with multiple techniques. Our tool has originally been developed for campaign data. Even datasets with little ancillary measurements or limited knowledge of the driving processes may be gap-filled. The only mandatory ancillary variable is a measurement of incoming radiation in order to distinguish between daytime and nighttime fluxes. If not available, the potential radiation (easily calculated from latitude and longitude) may be used instead. To setup the trace gas specific parameters, a description of the flux data and file properties needs to be provided in a configuration file (see example ini-file for tN_r in Table A.1).

The following gap-filling techniques have been implemented: Simple interpolation methods (Section 2.3.1), diurnal interpolation methods (Section 2.3.2), and look-up tables (Section 2.3.3). The gaps in the fluxes are filled following the two artificial gap-filling scheme 'hhs' and 'days' (Section 2.4) and the results are saved to ascii-files. Afterwards, the bootstrapping analysis is performed on the gap-filled fluxes. External results of gap-filled fluxes may be added to the analysis if the same artificial gap-filling scheme was used (as done for the artificial neural networks and inferential model in the manuscript). All the plots presented in this manuscript or supplements are automatically generated and the statistics and aggregated fluxes saved to ascii-files.

The original code of the multiple gap-filling tool developed for this manuscript with the tNr flux measurements as an example dataset are publicly available via the ICOS CarbonPortal DOI minting service: https://doi.org/10.18160/R6S5-47J1. The code is programmed in Python 3 and can be executed in a script or interactively accessed via a Jupyter Notebook. We are currently working on a user interface to easily incorporate new datasets and to use the MGF-Tool online at the Jupyter Hub of the Carbon Portal (https://jupyter3.icos-cp.eu/).

Table A.1
Settings in the ini-File for tN_r.

Variable	Settings	Description
Flux measurement settings		
FluxGas	"tNr"	Name of the trace gas
FluxColumn	"Nr_F"	Name of the flux column
FluxUnit	"ng N m-2 s-1"	Unit of the half-hourly fluxes
FluxFlag	"Nr_qc"	Name of the quality column (or "none")
FlagMax	1	Highest flag to be included (or nan)
ConvFactor	0.0180	Conversion factor from flux rate (FluxUnit) to half-hourly sum (ConvUnit)
ConvUnit	"g N ha-1 hh-1"	Unit of the aggregated fluxes
ConvSums	"g N ha-1 period-1"	Unit of the total period
Light measurement for daytime/nighttime differentiation		
LightVar	"Rg"	Column name of incoming radiation measurement
LightThres	5.0	Threshold of radiation for daytime/nighttime
Settings of the look-up table		
LUTVar_1	"Nr_C"	1st LUT variable
LUTRange_1	0.6	1st LUT variable range
LUTVar_2	"Rg"	2nd LUT variable
LUTRange_2	50	2nd LUT variable range
LUTVar_3	"ustar"	3rd LUT variable
LUTRange_3	0.1	3rd LUT variable range
Settings of a second LUT (used for MDS settings in the manuscript)		
LUT2Var_1	"Rg"	1st LUT2 variable
LUT2Range_1	50	1st LUT2 variable range
LUT2Var_2	"Temp"	2nd LUT2 variable
LUT2Range_2	2.5	2nd LUT2 variable range
LUT2Var_3	"VPD"	3rd LUT2 variable
LUT2Range_3	5.0	3rd LUT2 variable range
Additional setting for aggregating the fluxes		
DefGFT	"MDA_hh5_hhs"	Default alternative technique if a half-hour could not be filled by a gap-filling technique
Data file(s) settings		
FileData	"tNr_BaF_v07_mgf.csv"	Name of the ascii file with the original data
FileModels	"DE-BaF_NrX_ANN _v20191019.csv"	Name of the ascii file with additional gap-filled results
FileSeparator	" "	Value separator (delimiter) in ascii file
FileTimeStamp	"DateTime"	Name of the time stamp column

(Original naming of the variables can be used in the tool. If a variable name/value is not available, the setting needs to be set to "none"/nan. The ascii file should have Year-Month-Day as the order of the date format and period "." as a separator for decimal numbers).

Appendix A.2: Revisiting reported annual sums of NEE

A.2.1 Differences between annual sum estimates and error bounds

To compare differences of annual sum estimates of gap-filled fluxes with the error bounds of the used gap-filling techniques, the pioneering paper by Falge et al. (2001) and a recent publication by Vitale et al. (2019b) will be discussed as examples with typical results reported for CO_2. Both papers evaluate the errors of different gap-filling techniques and also show the results of annual sum NEE estimated for several sites.

In Falge et al. (2001), the absolute errors of the gap-filling technique were estimated using artificial gaps and ranged from 7 to 41 gC m^{-2} y^{-1} for three techniques. The absolute difference between the lowest and the highest estimate of the annual sum by the gap-filling techniques, *Delta* (Eq. 15), will serve as a measure of the discrepancy between annual sum estimates. The *Delta* between the techniques spanned from 5 to 201 gC m^{-2} y^{-1} for the 28 site years. In 21 cases, the annual sum estimates of at least one technique was outside the error bounds of the other techniques, including 13 cases where the error bounds were not even overlapping (Table A.2.1).

A comparison of two techniques with a very thorough analysis of the associated uncertainties by statistical inference can be found in Vitale et al. (2019b). The error bounds ranged from 13.2 to 137.6 gC m^{-2} y^{-1} and the *Delta* between techniques from 2 to 113 gC m^{-2} y^{-1} for the 10 site years. Despite the similar looking ranges, the annual sum estimates were outside the error bounds of the other technique in 6 cases (Table A.2.2), including 2 cases where the confidence intervals of the two techniques were not even overlapping.

The error bounds of the gap-filling techniques often underestimate the discrepancies found for the annual sum estimates between techniques. The total differences (*Deltas*) are the effect of the performance of the gap-filling technique *plus* the specific characteristics of the fluxes and real gaps in the dataset. Just to mention one example, the effect of a longer gap at the onset of the growing phase can only partly be simulated by inserting artificial gaps during the rest of the year or be captured by statistical inference.

A.2.2 Differences between sites and between techniques

The basic assumption behind using *one* standardized gap-filling technique is that this improves the inter-site (and inter-year) comparability of the predicted annuals sums. For example, if one gap-filling technique tends to overestimate the fluxes but always overestimates in a similar way, the absolute values of the annual sums would be affected the same way and hence be more comparable in relative terms. More generally, the differences of the annual sum estimates are expected to be smaller when using one technique across sites. This assumption is difficult to prove since even artificial gap scenarios only test a limited amount cases and the truth for the real gaps is unknown.

In the gap-filling comparison by Moffat et al. (2007) fifteen different gap-filling techniques were used and the median of the annual sum estimate of all techniques will be assumed to be close to the truth in the following. Their Table 5 states the deviations from the median for each technique and site year. To compare only "well-working" gap-filling techniques, only the deviations of the ten techniques with "medium" and "good" annual sum performance (see Table 3 therein) were taken and outliers removed, see Table A.2.3. The ten flux datasets had similar dataset characteristics in the sense that the eddy covariance measurements were all taken at forested European sites equipped with similar sensors and of high quality data including an extensive set of pre-filled

Table A.2.1
Average annual sums of net ecosystem exchange (NEE, gC m^{-2} y^{-1}) for three techniques with estimates of their absolute error taken from Table 6 in Falge et al. (2001) and their minima, maxima, and *Deltas*. The values are printed in bold if the absolute errors are smaller than the *Delta*. Additionally, the values are marked with an asterisk if the bounds of the absolute errors are non-overlapping.

Technique	MDC	Absolute error	LUT	Absolute error	NLR	Absolute error	*Min*	*Max*	*Delta*
Site Year									
WE97	177	**37**	114	**23**	104	**23**	*104*	*177*	***73****
TH97	-605	22	-603	15	-608	15	*-608*	*-603*	*5*
VI97	-328	**26**	-359	**27**	-368	**27**	*-368*	*-328*	***40***
LO97	-358	22	-363	14	-358	14	*-363*	*-358*	*5*
SO97	3	6	0	5	-3	5	*-3*	*3*	*6*
HY97	-272	17	-260	10	-266	10	*-272*	*-260*	*12*
HE97	-144	**7**	-153	**7**	-158	**7**	*-158*	*-144*	***14***
BR97	127	**41**	74	**26**	-74	**26**	*-74*	*127*	***201****
AB97	-614	**17**	-604	**11**	-624	**11**	*-624*	*-604*	***20***
WB95	-544	**16**	-517	**15**	-519	**15**	*-544*	*-517*	***27***
WB96	-796	**16**	-734	**15**	-738	**15**	*-796*	*-734*	***62****
WB97	-791	**19**	-698	**18**	-721	**18**	*-791*	*-698*	***93****
HL96	-321	**29**	-258	**18**	-278	**18**	*-321*	*-258*	***63****
HV92	-189	**28**	-324	**25**	-338	**25**	*-338*	*-189*	***149****
HV93	-210	31	-228	28	-225	28	*-228*	*-210*	*18*
HV94	-175	**14**	-162	**12**	-158	**12**	*-175*	*-158*	***17***
HV95	-230	17	-227	15	-229	15	*-230*	*-227*	*3*
HV96	-191	**12**	-170	**11**	-172	**11**	*-191*	*-170*	***21***
LW97	150	**6**	131	**10**	148	**10**	*131*	*150*	***19****
LW98	521	**7**	436	**13**	467	**13**	*436*	*521*	***85****
BV97	-563	**20**	-543	**16**	-526	**16**	*-563*	*-526*	***37****
BV98	125	**25**	133	**21**	165	**21**	*125*	*165*	***40***
SH97	-383	**8**	-349	**14**	-355	**14**	*-383*	*-349*	***34****
PO97	-147	41	-155	30	-174	30	*-174*	*-147*	*27*
ME96	-308	**16**	-287	**11**	-325	**11**	*-325*	*-287*	***38****
ME97	-328	**26**	-264	**17**	-324	**17**	*-328*	*-264*	***64****
DU98	-566	**40**	-555	**36**	-585	**36**	*-585*	*-555*	***30***
DU99	-708	**27**	-649	**25**	-666	**25**	*-708*	*-649*	***59****

(MDC – mean diurnal course, see Section 2.3.2, LUT – Look-up-table, see Section 2.3.3, NLR – nonlinear regression. More details on the table and techniques can be found in the original paper).

meteorological data.

The range between the lowest and the highest deviation of the annual sum estimate from the median will be called *Delta*. This *Delta* calculated from the median is essentially the same as the absolute difference between the lowest and highest annual sum prediction in Eq. 15. The *Delta* for the different techniques ($Delta_{Techniques}$) ranged from 12.8 to 45.1 g C m^{-2} y^{-1} over all site-years and the *Delta* between the different site years ($Delta_{Sites}$) from 14.1 to 42.6 g C m^{-2} y^{-1} over all techniques (Table A.2.3). Hence, using the same technique leads to very similar gap-filling errors on the annual sums among all site-years (vertical) as using all well-working techniques for the same site-year (horizontal). Furthermore, the *Deltas* of the same technique at the same site but different years can be as small as 2.0 g C m^{-2} y^{-1} but also range up to 45.1 g C m^{-2} y^{-1}. Two techniques are having their highest *Delta* for two different years of the same site.

These results for multiple techniques and multiple sites show no indication that using only one technique would reduce the gap-filling error on the annual sum estimate. Annual sum estimates may be just as comparable if different well-working gap-filling techniques are used.

Table A.2.2
Annual budget estimates for net ecosystem exchange (NEE, gC m^{-2} y^{-1}) for two techniques with lower and upper confidence intervals (CI) and total uncertainty taken from Table 2 from Vitale et al. (2019). and their *Deltas*. The values are printed in bold if the total uncertainties are smaller than the *Delta*. Additionally, the values are marked with an asterisk if the confidence intervals are non-overlapping.

Technique	Lower CI	Budget MDS	Upper CI	Total uncertainty	Lower CI	Budget PADL	Upper CI	Total uncertainty	*Delta*
Site Year									
AT-Neu 2010	509	558	608	**49.4**	556	645	733	**88.4**	***87***
AU-Cpr 2012	-216	-203	-190	**13.2**	-250	-232	-214	**18.2**	***29***
AU-How 2011	-618	-576	-533	**42.4***	-756	-689	-621	**67.6***	***113****
DK-Sor 2009	-359	-314	-268	45.4	-360	-312	-264	47.9	*2*
FI-Hyy 2007	-264	-240	-216	**24.3**	-314	-282	-251	**31.5**	***42***
FR-Pue 2008	-309	-285	-261	**23.9***	-387	-355	-322	**32.6***	***70****
GF-Guy 2008	-172	-103	-34	68.8	-223	-85	52	137.6	*18*
IT-CA1 2012	-349	-319	-290	**29.6**	-423	-383	-344	**39.5**	***64***
US-Los 2006	-34	-13	8	21.2	-57	-15	27	41.8	*2*
US-Ne2 2012	-537	-452	-368	84.3	-566	-480	-395	85.9	*28*

(PADS - panel autoregressive distributed lag multiple imputation model, MDS - marginal distribution sampling. More details on the table and techniques can be found in the original paper.)

Table A.2.3
Deviations from the median over all techniques of the annual sum predictions of net ecosystem exchange (NEE , gC m^{-2} y^{-1}) taken from Table 5 in Moffat et al. (2007) and their horizontal and vertical minima, maxima, and *Deltas*. The rows of the two techniques having the highest *Delta* at the same site for two different years are printed in bold.

Technique	NLR _AM	NLR FCRN _STD	NLR _FM _OLS	NLR _LM	ANN _BR	ANN _PS	LUT	MDS	SPM	MDV	*Min*	*Max*	$Delta_{Sites}$
Site Year													
be1_2000	4.4	-1.6	-2.4	10.6	-13	0	10	**-12.7**	1.9	2.5	*-13*	*10.6*	*23.6*
be1_2001	0.3	-1.1	-6.7	7.4	-2	2.7	5.9	**4.8**	0	-	*-6.7*	*7.4*	*14.1*
de3_2000	15.7	2.4	**-21.7**	7.3	-	2	0	-10.5	-7.9	-0.9	*-21.7*	*15.7*	*37.4*
de3_2001	-	5.3	**23.4**	-16.9	-6.1	0	-19.2	2.9	-6.1	9.9	*-19.2*	*23.4*	*42.6*
fi1_2001	-8	0.4	-1.6	0	-9.6	-4.6	1	1.9	-	8.8	*-9.6*	*8.8*	*18.4*
fi1_2002	6.3	0	-2.3	2.9	-12	2.2	-2.5	-6.7	19.9	6.5	*-12*	*19.9*	*31.9*
fr1_2001	7.4	-2.1	-0.7	2	8.9	-2.8	0	4.6	15.6	-12.1	*-12.1*	*15.6*	*27.7*
fr1_2002	-3.8	18.2	-6.5	0	-2	-10.1	5.3	1.4	1.9	-10.8	*-10.8*	*18.2*	*29*
fr4_2002	14.3	-7.4	18.2	6.4	-	-1.7	1.6	-0.5	-7	0	*-7.4*	*18.2*	*25.6*
it3_2002	9.6	0	6	21.6	-6.5	-0.6	8.4	-6.7	7.6	3.7	*-6.7*	*21.6*	*28.3*
Min	*-8*	*-7.4*	*-21.7*	*-16.9*	*-13*	*-10.1*	*-19.2*	*-12.7*	*-7.9*	*-12.1*			
Max	*15.7*	*18.2*	*23.4*	*21.6*	*8.9*	*2.7*	*10*	*4.8*	*19.9*	*9.9*			
$Delta_{Techniques}$	*23.7*	*25.6*	*45.1*	*38.5*	*21.9*	*12.8*	*29.2*	*17.5*	*27.8*	*22*			

(More details on the table and techniques can be found in the original paper.)

References

Ammann, C., Wolff, V., Marx, O., Brümmer, C., Neftel, A., 2012. Measuring the biosphere-atmosphere exchange of total reactive nitrogen by eddy covariance. Biogeosciences 9 (11), 4247–4261.

Aubinet, M., Vesala, T., Papale, D.E., 2012. Eddy Covariance – A Practical Guide to Measurement and Data Analysis. Springer, Dordrecht, Heidelberg, London, New York.

Baldocchi, D., et al., 2001. FLUXNET: A New Tool to Study the Temporal and Spatial Variability of Ecosystem–Scale Carbon Dioxide, Water Vapor, and Energy Flux Densities. Bulletin of the American Meteorological Society 82 (11), 2415–2434.

Baldocchi, D.D., 2019. How eddy covariance flux measurements have contributed to our understanding of Global Change Biology. Global Change Biology 26 (1), 242–260.

Brown, M., et al., 2010. Impact of mountain pine beetle on the net ecosystem production of lodgepole pine stands in British Columbia. Agricultural and Forest Meteorology 150 (2), 254–264.

Brümmer, C., et al., 2012. How climate and vegetation type influence evapotranspiration and water use efficiency in Canadian forest, peatland and grassland ecosystems. Agricultural and Forest Meteorology 153, 14–30.

Brümmer, C., et al., 2008. Diurnal, seasonal, and interannual variation in carbon dioxide and energy exchange in shrub savanna in Burkina Faso (West Africa). Journal of Geophysical Research: Biogeosciences 113 (G2), 1–11.

Brümmer, C., et al., 2013. Fluxes of total reactive atmospheric nitrogen (ΣNr) using eddy covariance above arable land. Tellus B 65 (19770), 1–20.

Denmead, O.T., et al., 2010. Emissions of methane and nitrous oxide from Australian sugarcane soils. Agricultural and Forest Meteorology 150 (6), 748–756.

Drought. Team and ICOS Ecosystem Thematic Centre, 2020. Drought-2018 ecosystem eddy covariance flux product for 52 stations in FLUXNET-Archive format. https://doi.org/10.18160/YVR0-4898.

Efron, B., Tibshirani, R.J., 1994. An Introduction to the Bootstrap. Taylor & Francis.

Falge, E., et al., 2001. Gap filling strategies for defensible annual sums of net ecosystem exchange. Agricultural and Forest Meteorology 107, 43–69.

Famulari, D., et al., 2004. Measuring Eddy Covariance Fluxes of Ammonia Using Tunable Diode Laser Absorption Spectroscopy. Water, Air, & Soil Pollution: Focus 4 (6), 151–158.

Fernández-Martínez, M., et al., 2014. Nutrient availability as the key regulator of global forest carbon balance. Nat. Clim. Chang. 4 (6), 471–476.

Flechard, C.R., et al., 2011. Dry deposition of reactive nitrogen to European ecosystems: a comparison of inferential models across the NitroEurope network. Atmospheric Chemistry and Physics 11 (6), 2703–2728.

Fleischer, K., et al., 2013. The contribution of nitrogen deposition to the photosynthetic capacity of forests. Global Biogeochemical Cycles 27 (1), 187–199.

Graf, A., et al., 2020. Altered energy partitioning across terrestrial ecosystems in the European drought year 2018. Philos Trans R Soc Lond B Biol Sci 375 (1810), 20190524. -20190524.

Heiskanen, J., et al., 2022. The Integrated Carbon Observation System in Europe. Bulletin of the American Meteorological Society 103 (3), E855–E872.

Herbst, M., Friborg, T., Ringgaard, R., Soegaard, H., 2011. Interpreting the variations in atmospheric methane fluxes observed above a restored wetland. Agricultural and Forest Meteorology 151 (7), 841–853.

Hollinger, D.Y., Richardson, A.D., 2005. Uncertainty in eddy covariance measurements and its application to physiological models. Tree Physiology 25 (7), 873–885.

Horii, C., et al., 2006. Atmospheric reactive nitrogen concentration and flux budgets at a Northeastern U.S. forest site. Agricultural and Forest Meteorology 136 (3-4), 159–174.

Horii, C.V., Munger, J.W., Wofsy, S.C., 2004. Fluxes of nitrogen oxides over a temperate deciduous forest. Journal of Geophysical Research 109 (D8), 1–8.

Hurkuck, M., Brümmer, C., Kutsch, W.L., 2016. Near-neutral carbon dioxide balance at a seminatural, temperate bog ecosystem. Journal of Geophysical Research: Biogeosciences 121 (2), 370–384.

Hurkuck, M., et al., 2014. Determination of atmospheric nitrogen deposition to a semi-natural peat bog site in an intensively managed agricultural landscape. Atmos. Environ. 97, 296–309.

Irvin, J., et al., 2021. Gap-filling eddy covariance methane fluxes: Comparison of machine learning model predictions and uncertainties at FLUXNET-CH4 wetlands. Agricultural and Forest Meteorology 308-309, 1–22.

Keenan, T.F., et al., 2014. Net carbon uptake has increased through warming-induced changes in temperate forest phenology. Nat. Clim. Chang. 4 (7), 598–604.

Keenan, T.F., et al., 2013. Increase in forest water-use efficiency as atmospheric carbon dioxide concentrations rise. Nature 499 (7458), 324–327.

Kim, Y., et al., 2019. Gap-filling approaches for eddy covariance methane fluxes: A comparison of three machine learning algorithms and a traditional method with principal component analysis. Global Change Biology 160 (12), 12.

Knohl, A., Schulze, E.-D., Kolle, O., Buchmann, N., 2003. Large carbon uptake by an unmanaged 250-year-old deciduous forest in Central Germany. Agricultural and Forest Meteorology 118 (3-4), 151–167.

Knox, S.H., et al., 2019. FLUXNET-CH4 Synthesis Activity: Objectives, Observations, and Future Directions. Bulletin of the American Meteorological Society 100 (12), 2607–2632.

Kutsch, W.L., et al., 2008. Advection and resulting CO2 exchange uncertainty in a tall forest in Central Germany. Ecological Applications 18 (6), 1391–1405.

Lasslop, G., Reichstein, M., Kattge, J., Papale, D., 2008. Influences of observation errors in eddy flux data on inverse model parameter estimation. Biogeosciences 5, 1311–1324.

Lindauer, M., et al., 2014. Net ecosystem exchange over a non-cleared wind-throw-disturbed upland spruce forest—Measurements and simulations. Agricultural and Forest Meteorology 197, 219–234.

Magnani, F., et al., 2007. The human footprint in the carbon cycle of temperate and boreal forests. Nature 447 (7146), 849–851.

Mahabbati, A., et al., 2021. A comparison of gap-filling algorithms for eddy covariance fluxes and their drivers. Geosci. Instrum. Method. Data Syst. 10 (1), 123–140.

Marx, O., Brümmer, C., Ammann, C., Wolff, V., Freibauer, A., 2012. TRANC - a novel fast-response converter to measure total reactive atmospheric nitrogen. Atmospheric Measurement Techniques 5 (5), 1045–1057.

Massad, R.S., Nemitz, E., Sutton, M.A., 2010. Review and parameterisation of bi-directional ammonia exchange between vegetation and the atmosphere. Atmospheric Chemistry and Physics 10 (21), 10359–10386.

Mauder, M., Foken, T., 2006. Impact of post-field data processing on eddy covariance flux estimates and energy balance closure. Meteorologische Zeitschrift 15 (6), 597–609.

Menzer, O., et al., 2013. Random errors in carbon and water vapor fluxes assessed with Gaussian Processes. Agricultural and Forest Meteorology 178–179, 161–172.

Metzger, S., et al., 2019. From NEON Field Sites to Data Portal: A Community Resource for Surface–Atmosphere Research Comes Online. Bulletin of the American Meteorological Society 100 (11), 2305–2325.

Moffat, A.M., 2012. A new methodology to interpret high resolution measurements of net carbon fluxes between terrestrial ecosystems and the atmosphere. Friedrich Schiller University, Jena, Germany. Doctoral Thesis. http://www.db-thueringen.de/servlets/DocumentServlet?id=20321.

Moffat, A.M., Beckstein, C., Churkina, G., Mund, M., Heimann, M., 2010. Characterization of ecosystem responses to climatic controls using artificial neural networks. Global Change Biology 16, 2737–2749.

Moffat, A.M., et al., 2007. Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. Agricultural and Forest Meteorology 147, 209–232.

Neftel, A., et al., 2010. N2O exchange over managed grassland: Application of a quantum cascade laser spectrometer for micrometeorological flux measurements. Agricultural and Forest Meteorology 150 (6), 775–785.

Nemitz, E., et al., 2009. Intercomparison and assessment of turbulent and physiological exchange parameters of grassland. Biogeosciences 6, 1445–1466.

Nemitz, E., et al., 2018. Standardisation of eddy-covariance flux measurements of methane and nitrous oxide. International Agrophysics 32 (4), 517–549.

Nemitz, E., Milford, C., Sutton, M.A., 2001. A two-layer canopy compensation point model for describing bi-directional biosphere-atmosphere exchange of ammonia. Quarterly Journal of the Royal Meteorological Society 127 (573), 815–833.
Odum, E.P., 1969. The Strategy of Ecosystem Development. Science 164 (3877), 262–270.
Pastorello, G., et al., 2020. The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. Scientific Data 7 (1), 1–27.
Reichstein, M., et al., 2005. On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. Global Change Biology 11 (9), 1424–1439.
Richardson, A.D., et al., 2012. Uncertainty quantification. In: Aubinet, M., Vesala, T., Papale, D. (Eds.), Eddy covariance. Springer, Dordrecht, pp. 173–209.
Richardson, A.D., Hollinger, D.Y., 2007. A method to estimate the additional uncertainty in gap-filled NEE resulting from long gaps in the CO_2 flux record. Agricultural and Forest Meteorology 147, 199–208.
Richardson, A.D., et al., 2008. Statistical properties of random CO2 flux measurement uncertainty inferred from model residuals. Agricultural and Forest Meteorology 148 (1), 38–50.
Schrader, F., et al., 2016. Non-stomatal exchange in ammonia dry deposition models: comparison of two state-of-the-art approaches. Atmos. Chem. Phys. 16 (21), 13417–13430.
Schrader, F., Erisman, J.W., Brümmer, C., 2020. Towards a coupled paradigm of NH 3 -CO 2 biosphere–atmosphere exchange modelling. Global Change Biology 26 (9), 4654–4663.
Sintermann, J., et al., 2011. Eddy covariance flux measurements of ammonia by high temperature chemical ionisation mass spectrometry. Atmospheric Measurement Techniques 4 (3), 599–616.
Tang, A.C.I., et al., 2018. Eddy Covariance Measurements of Methane Flux at a Tropical Peat Forest in Sarawak, Malaysian Borneo. Geophysical Research Letters 45 (9), 4390–4399.
van Dijk, A.I.J.M., Dolman, A.J., Schulze, E.-D., 2005. Radiation, temperature, and leaf area explain ecosystem carbon fluxes in boreal and temperate European forests. Global Biogeochemical Cycles 19 (2) n/a-n/a.
Vernadsky, V.I., 1998. The Biosphere. Springer-Verlag, New York.
Vitale, D., Bilancia, M., Papale, D., 2019a. Modelling random uncertainty of eddy covariance flux measurements. Stochastic Environmental Research and Risk Assessment 33 (3), 725–746.
Vitale, D., Bilancia, M., Papale, D., 2019b. A Multiple Imputation Strategy for Eddy Covariance Data. J ENV INFORM 34 (2), 68–87.
Wallach, D., Mearns, L.O., Ruane, A.C., Rötter, R.P., Asseng, S., 2016. Lessons from climate modeling on the design and use of ensembles for crop modeling. Climatic Change 139 (3), 551–564.
Wang, H.-J., Riley, W.J., Collins, W.D., 2015. Statistical uncertainty of eddy covariance CO2 fluxes inferred using a residual bootstrap approach. Agricultural and Forest Meteorology 206, 163–171.
Wesely, M.L., 1989. Parameterization of surface resistances to gaseous dry deposition in regional-scale numerical models. Atmos. Environ. 23 (6), 1293–1304.
Wintjen, P., Ammann, C., Schrader, F., Brümmer, C., 2020. Correcting high-frequency losses of reactive nitrogen flux measurements. Atmospheric Measurement Techniques 13 (6), 2923–2948.
Wintjen, P., Schrader, F., Schaap, M., Beudert, B., Brümmer, C., 2022. Forest–atmosphere exchange of reactive nitrogen in a remote region – Part I: Measuring temporal dynamics. Biogeosciences 19, 389–413.
Wutzler, T., et al., 2018. Basic and extensible post-processing of eddy covariance flux data with REddyProc. Biogeosciences 15 (16), 5015–5030.
Zhu, S., Clement, R., McCalmont, J., Davies, C.A., Hill, T., 2022. Stable gap-filling for longer eddy covariance data gaps: A globally validated machine-learning approach for carbon dioxide, water, and energy fluxes. Agricultural and Forest Meteorology 314, 1–10.
Zöll, U., et al., 2016. Surface–atmosphere exchange of ammonia over peatland using QCL-based eddy-covariance measurements and inferential modeling. Atmospheric Chemistry and Physics 16 (17), 11283–11299.
Zöll, U., et al., 2019. Is the biosphere-atmosphere exchange of total reactive nitrogen above forest driven by the same factors as carbon dioxide? An analysis using articial neural networks. Atmos. Environ. 206, 108–118.