

1 **Stable gap-filling for longer eddy covariance data gaps: a globally validated** 2 **machine-learning approach for carbon dioxide, water, and energy fluxes**

3 **Songyan Zhu¹, Robert Clement¹, Jon McCalmont¹, Christian A. Davies², Timothy Hill¹**

4 ¹ College of Life and Environmental Science, University of Exeter, Streatham Campus, Rennes Drive. Exeter, EX4 4RJ. UK

5 ² Shell International Exploration and Production Inc., Shell Technology Centre Houston, Houston, TX 77082, USA

6 **Abstract**

7 Continuous time-series of CO₂, water, and energy fluxes are useful for evaluating the impacts
8 of climate-change and management on ecosystems. The eddy covariance (EC) technique can
9 provide continuous, direct measurements of ecosystem fluxes, but to achieve this gaps in data
10 must be filled. Research-standard methods of gap-filling fluxes have tended to focus on CO₂ fluxes
11 in temperate forests and relatively short gaps of less than two weeks. A gap-filling method
12 applicable to other fluxes and capable of filling longer gaps is needed.

13 To address this challenge, we propose a novel gap-filling approach, Random Forest Robust
14 (RFR). RFR can accommodate a wide range of data gap sizes, multiple flux types (i.e. CO₂, water
15 and energy fluxes). We configured RFR using either three (RFR₃) or ten (RFR₁₀) driving variables.
16 RFR was tested globally on fluxes of CO₂, latent heat (LE), and sensible heat (H) from 94 suitable
17 FLUXNET2015 sites by using artificial gaps (from 1 to 30 days in length) and benchmarked against
18 the standard marginal distribution sampling (MDS) method.

19 In general, RFR improved on MDS's R² by 15 % (RFR₃) and by 30 % (RFR₁₀) and reduced
20 uncertainty by 70 %. RFR's improvements in R² for H and LE were more than twice the
21 improvement observed for CO₂ fluxes. Unlike MDS, RFR performed well for longer gaps; for
22 example, the R² of RFR methods in filling 30-day gaps dropped less than 4 % relative to 1-day gaps,
23 while the R² of MDS dropped by 21 %.

24 Our results indicate that the RFR method can provide improved gap-filling of CO₂, H and LE
25 flux timeseries. Such improved continuous flux measurements, with low bias, can enhance our
26 understanding of the impacts of climate-change and management on ecosystems globally.

27

28 Keywords: global land ecosystems, carbon exchange, eddy covariance, long gaps, robust gap-
29 filling

30 1. Introduction

31 To keep climate change to below 1.5°C within reach (Wollenberg et al. 2016; Glanemann et al.
32 2020; Smith et al. 2021), Natural Climate Solutions (NCS) (Griscom et al. 2017) may be the most
33 cost-effective approach immediately ready for large-scale deployment (Cohen-Shacham et al.
34 2019), because land ecosystems absorb approximately one third of anthropogenic C emission per
35 year (Friedlingstein et al. 2020). NCS have already been implemented in 66 % of countries
36 (Chausson et al. 2020), but measuring and verifying the effectiveness of NCS remains challenging
37 (Skinner and Dell 2015; Smith et al. 2020; Bautista et al. 2021).

38 Eddy covariance (EC) has been suggested as part of the solution to the NCS measurement
39 challenge e.g. inaccessible and hard-to-observe carbon pool changes (Baldocchi 2020; Keith et al.
40 2021; Hemes et al. 2021). EC can monitor (ecosystem-scale) mass (CO₂, water, CH₄, and N₂O.) and
41 energy fluxes continuously (Aubinet et al. 2012; Hill et al. 2017; Baldocchi 2020), with a broad
42 convergence between EC and other carbon exchange quantification methods (Skinner and Dell
43 2015; Campioli et al. 2016). Currently over 400 EC towers are contributing datasets to the global
44 synthesis project FLUXNET (Baldocchi et al. 2001; Baldocchi 2014; Pastorello et al. 2020).

45 However, data gaps hinder the application of EC flux time-series (Aubinet et al. 2012). Most
46 EC data gaps occur as a result of instrument failure (e.g. power loss and sensor malfunction)
47 (Papale et al. 2006), rejection of data during quality control (Mauder et al. 2008), and data loss
48 through adverse environmental conditions (Falge et al. 2001). Gap-filling approaches for EC
49 include the research-standard Marginal Distribution Sampling (MDS) (Reichstein et al. 2005;

50 Pastorello et al. 2020), which fills gaps by considering the covariance of fluxes with meteorological
51 drivers (global radiation, air temperature and vapour pressure deficit) and the temporal
52 autocorrelation of the flux values (Reichstein et al. 2005), and other numerical methods (e.g.
53 machine-learning) aiming for improving gap-filling performance (Vitale et al. 2019; Irvin et al.
54 2021).

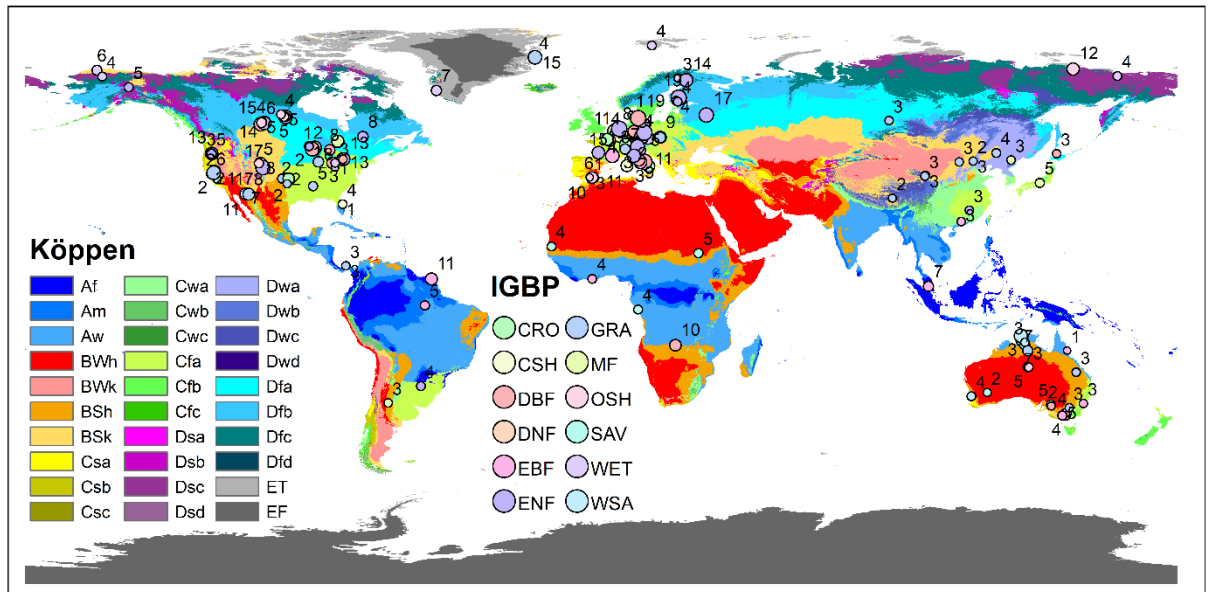
55 Previous comparisons of gap-filling approaches have tended to focus on gaps in carbon fluxes
56 of up to two weeks in temperate forests (Moffat et al. 2007) despite being routinely applied
57 globally to carbon, water, and heat fluxes. Whilst MDS has been demonstrated as an effective gap-
58 filling method for filling short gaps using a small driver set (Moffat et al. 2007), it was reported not
59 designed for long temporal data gaps (Kang et al. 2019). Additional uncertainty in filled long NEE
60 gaps (~ three weeks) was reported (Richardson and Hollinger 2007), but still no robust methods
61 have been proposed for filling long gaps. Machine-learning (e.g. random forest) methods
62 outperformed MDS in filling, e.g. methane flux, gaps, but they require 7-14 drivers (e.g. leaf area
63 index) to fill gaps (Menzer et al. 2015; Kim et al. 2020). It remains unknown if more recent
64 machine-learning methods can improve on MDS for the same driver sets and as machine learning
65 can leverage information from a larger, expanded driver set.

66 In this paper, we present a gap-filling approach for NEE (CO₂ fluxes), H (sensible heat), and LE
67 (latent energy), based on a new Random Forest-Robust (RFR) algorithm, that is designed to be
68 effective for longer data gaps. RFR was implemented using two different driver sets to simulate
69 good and poor driver availability: 1) the same three meteorological drivers as MDS (RFR₃) and 2)
70 an expansion to ten drivers (RFR₁₀) to explore if additional gap-filling improvements can be seen
71 by exploiting this wider range of drivers. We evaluated RFR₃ and RFR₁₀ against MDS by using 94
72 globally distributed sites (806 site-years) from the FLUXNET database. Gap-filling and validation
73 were carried out for artificial gaps much longer than previous validations (Moffat et al. 2007), with
74 a combination of short (24-hour), long (7-day), and very long (30-day) missing periods. Finally, we
75 independently verified gap-filling performance by comparing the EBR (energy balance ratio) of

76 measured data to the EBR of gap-filled data. To explore the limitations of approaches, gap-filling
 77 performance was examined for daytime and night-time periods and for different international
 78 Geosphere–Biosphere Programme (IGBP) ecosystems surface classifications.

79

80 2. Methodology



81 *Figure 1 FLUXNET2015 sites (dots) used for gap-filling. The underlying map represents the Köppen climate classifications.*
 82 *Dot colours represent the International Geosphere-Biosphere Programme (IGBP) land cover classification. Dot sizes*
 83 *represent the data length in years of sites (noted by the numbers aside).*

84

85 2.1. FLUXNET 2015 site selection

86 The FLUXNET 2015 dataset contains open access data (at half-hourly resolution) from 206
 87 globally distributed sites, comprising quality-controlled ecosystem-scale NEE, H, and LE fluxes
 88 along with associated meteorological and biological variables (Pastorello et al. 2020). Whilst
 89 installed and maintained by different researchers, a uniform flux post-processing procedure was
 90 applied to all sites (Pastorello et al. 2017, 2020). We used half-hourly FLUXNET 2015 products:
 91 NEE_VUT_REF, NEE_VUT_REF_QC, H_F_MDS, H_F_MDS_QC, LE_F_MDS, and LE_F_MDS_QC
 92 (<https://fluxnet.org/data/fluxnet2015-dataset/fullset-data-product/>). Quality control flags (*_QC)
 93 were used to identify gap-filled fluxes already present in the datasets. Not all 206 sites were

94 appropriate for validating gap-filling approaches (sites used and their background information are
95 shown in Figure 1 and Table S1-S2), 48 sites did not provide quality control information for H and
96 LE and 86 did not have the required drivers to implement RFR₁₀. In addition, 12 sites did not
97 contain enough original (non-gap-filled) data to accommodate the artificial gaps for validation.
98 Due to these constraints, a sub-set of 94 sites were analysed for gap-filling for the complete NEE,
99 H, and LE.

100 2.1.1. Environmental gap filling drivers

101 We used pre-filled environmental drivers provided by the FLUXNET2015 database. Drivers for
102 MDS and RFR₃ were downward shortwave radiation (SW_IN_F), vapour pressure deficit
103 (VPD_F_MDS), and air temperature (TA_F_MDS). The additional seven drivers for the extended
104 RFR₁₀ were net radiation (NETRAD), wind speed (WS), wind direction (WD), soil heat flux
105 (G_F_MDS), soil temperature (TS_F_MDS), relative humidity (RH), and soil water content
106 (SWC_F_MDS).

107 2.1.2. Site characteristic descriptors

108 For each site, we extracted descriptors of geographical location, land-use classification, local
109 meteorology, climate classification, and instrumental setup to provide comprehensive
110 information on gap-filling performance analysis (Table S1-S2). Descriptors extracted from the
111 FLUXNET site meta-data include continent, altitude, the International Geosphere-Biosphere
112 Programme (IGBP), and Koppen's climate classification (E Falge et al., 2017; Gilberto et al., 2020).
113 From the FLUXNET2015 database we extracted mean annual temperature (°C), precipitation (mm)
114 and wind speed (m s⁻¹). Instrumental setup was classified by sensor type (i.e. open-path, closed-
115 path, or both), instrument-to-canopy height ratio and data set duration. Information on site setup
116 was determined by a literature search of the primary publications for each site.

117

118

119 2.2. Artificial gap scenario

120 Artificial gaps were generated within the datasets to be filled using three approaches; 25 % of
121 total half-hours were randomly removed comprised of three different gap lengths: short gaps (24-
122 hour, 20 % of total gaps), long gaps (7-day, 30 % of total gaps) and very-long gaps (30-day, 50 %
123 of total gaps). Differently located random gap scenarios were generated for each site. For each
124 site, NEE, H, and LE shared the same gaps. Where the artificial gaps overlapped with existing 'real'
125 gaps we required at least 50 % original measured data be present, if this criterion was not met,
126 the artificial gap was discarded and randomly re-generated until it meets the '>50 %' criterion.
127 Sites with insufficient original measured data to provide the required gap lengths were rejected
128 from the analysis.

129

130 2.3. Gap-filling approaches

131 The benchmark MDS was implemented using the R package REddyProc (v. 1.2.2) (Wutzler et
132 al. 2018), further details on the MDS approach can be found in (Reichstein et al. 2005). Our novel
133 machine learning approach, Random Forest Robust (RFR), was developed using the 'fluxlib'
134 package (<https://github.com/soonyenju/fluxlib>) in Python 3.6+, and is based on Random Forest
135 implemented in Scikit-Learn (v. 0.24.1) (Pedregosa et al. 2011) with a new feature selector called
136 'receptive limiter' (details are given in section **Error! Reference source not found.**). Training of the
137 RFR was performed for each site separately. Because our RFR approach contains two distinct
138 driver sets, a total of three methods (MDS, RFR₃ and RFR₁₀) were validated at each site.

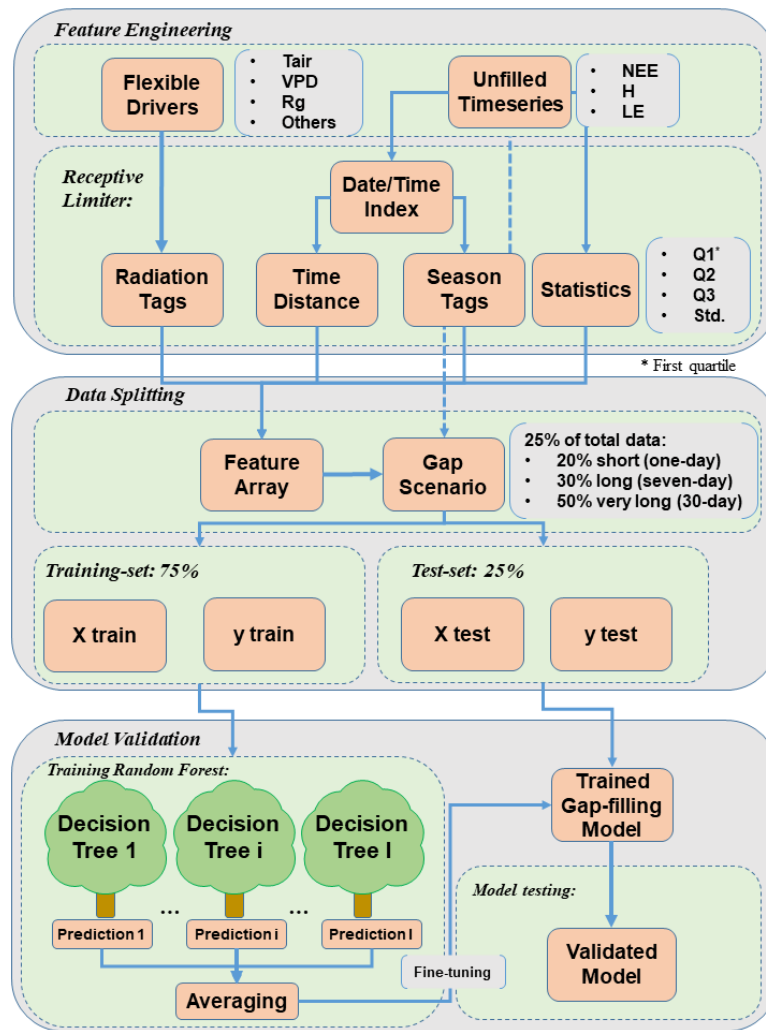


Figure 2. Workflow of implementing RFR: Feature engineering (top grey panel), data splitting via gap scenario (middle grey panel), and model validation (bottom grey panel).

139 The RFR approach has been widely implemented in ecological applications (Breiman 2001;
 140 Jung et al. 2009; Tramontana et al. 2015; Zeng et al. 2020). Our implementation of RFR comprises
 141 three steps: feature engineering, data splitting, and model validation (Figure 2).

142 The ‘Receptive Limiter’ is the core of feature engineering, continuous data are extracted and
 143 binned into discrete categories, downward solar radiation is tagged as, for example, ‘weak’ (< 10
 144 W m⁻²), ‘medium’ (10 – 100 W m⁻²), or ‘strong’ (> 100 W m⁻²). Time distance from the beginning of
 145 the time-series (in hours) is extracted as a feature to capture the potential ecosystem growing or
 146 degrading trends. Seasons (by the month of time-series) are tagged as ‘winter’ (DJF in North
 147 Hemisphere; JJA in South Hemisphere), ‘spring’ (MAM in North Hemisphere; SON in South

148 Hemisphere), 'summer' (JJA in North Hemisphere; DJF in South Hemisphere), and 'autumn' (SON
149 in North Hemisphere; MAM in South Hemisphere). Daily flux quartiles and standard deviations are
150 extracted from quality-controlled flux time-series as RFR input features separately from NEE, H,
151 and LE to preclude potential outliers in filled gaps. Features and fluxes are split into training and
152 testing data (training-set and test-set). Training data is used to separately feed the RFR.
153 Hyperparameters of RFR are automatically optimized using the GridSearchCV function of Scikit-
154 Learn. The trained RFR models are subsequently validated against the test-set.

155

156 2.4. Evaluation indicators

157 Statistical comparisons between gap-filled and original measured values within the artificial
158 gaps were carried out for NEE, H, and LE at each site using the coefficient of determination (R^2),
159 slope of linear regression, Root Mean Squared Error (RMSE, $\text{g C (carbon) m}^{-2} \text{ d}^{-1}$ for NEE and W m^{-2}
160 2 for H and LE), and bias (same units as RMSE).

161 The bias is defined as:

$$162 \quad \text{bias} = \frac{\sum \text{Fill.} - \sum \text{Meas.}}{n}$$

163 Where:

164 *Fill.* denotes the filled gaps

165 *Meas.* denotes the measured fluxes (of corresponding artificial gaps)

166 n is the length of gaps measured as the number of half-hours

167

168 These descriptive statistics are also determined separately for daytime and night-time periods,
169 where daytime is defined as periods above a threshold of $20 \text{ W m}^{-2} \text{ Rg}$ (Papale et al. 2006).

170 Welch's T-test (Derrick et al. 2016) was used to determine gap-filling improvement by RFR₃
171 over MDS and by RFR₁₀ over RFR₃ separately within the 95 % confidence interval.

172 We use the energy balance ratio (EBR) of the gap-filled periods as an independent measure of
 173 gap-filling bias in the energy fluxes (i.e. LE and H) (Foken et al. 2011; Perez-Priego et al. 2017).
 174 According to the following formula (Eshonkulov et al. 2019):

$$175 \quad EBR = \frac{\sum(H + LE)}{\sum(NETRAD - G)}$$

176
 177 Where:

178 EBR = energy balance ratio

179 $NETRAD$ = ground downward net radiation ($W m^{-2}$), derived from FLUXNET2015

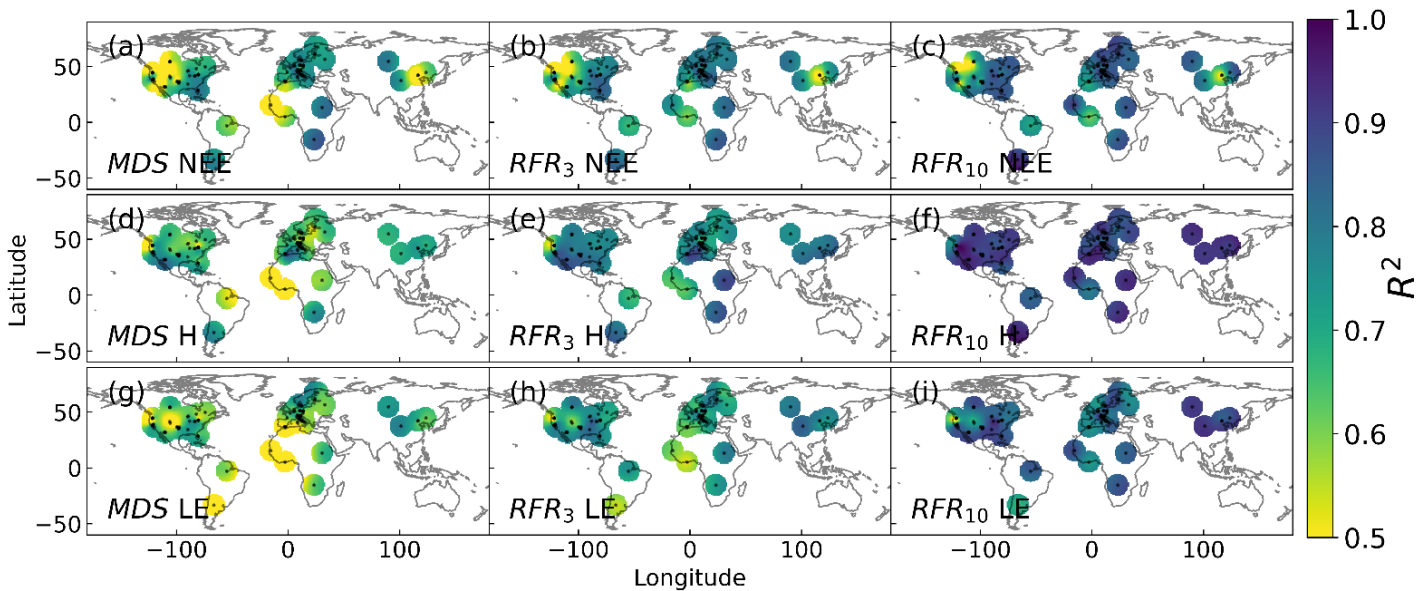
180 G = ground heat flux ($W m^{-2}$), derived from FLUXNET2015

181

182 3. Results

183 3.1. Gap-filling performance

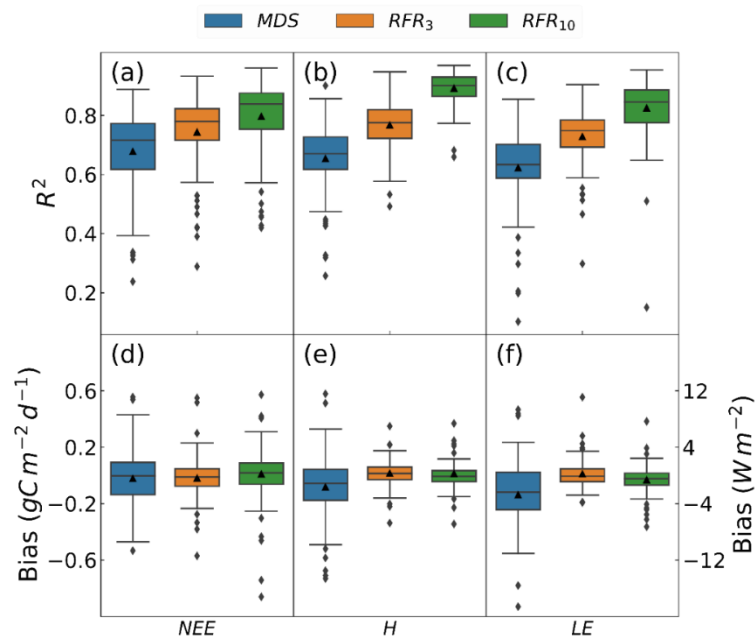
184



185 *Figure 3 R² map of comparing MDS, RFR₃, and RFR₁₀ filled gaps with measurements for NEE, H, and LE, respectively. It*
 186 *shows spatial distribution of the variance explained by gap-filling (R²) at 94 FLUXNET2015 sites. (We also provide validation*
 187 *at 194 sites (1346 site-years) covering six continents, 11 IGBP classes, and 18 Koppen climate classes (Table S9)).*

188

189 In general, North America and Europe comprised the most sites and Europe was seen with the
 190 highest R^2 for NEE, H, and LE; while South America and Africa were seen with the lowest for H and
 191 LE (Figure 3). Comparing NEE with H and LE, northwest North America and northeast Asia were
 192 seen with low R^2 ; but R^2 for NEE in South America and Africa were relatively higher. As regards to
 193 gap-filling approaches, RFR₃ was seen with higher R^2 over MDS, and RFR₁₀ was seen with further
 194 higher R^2 , especially in South America and Africa.

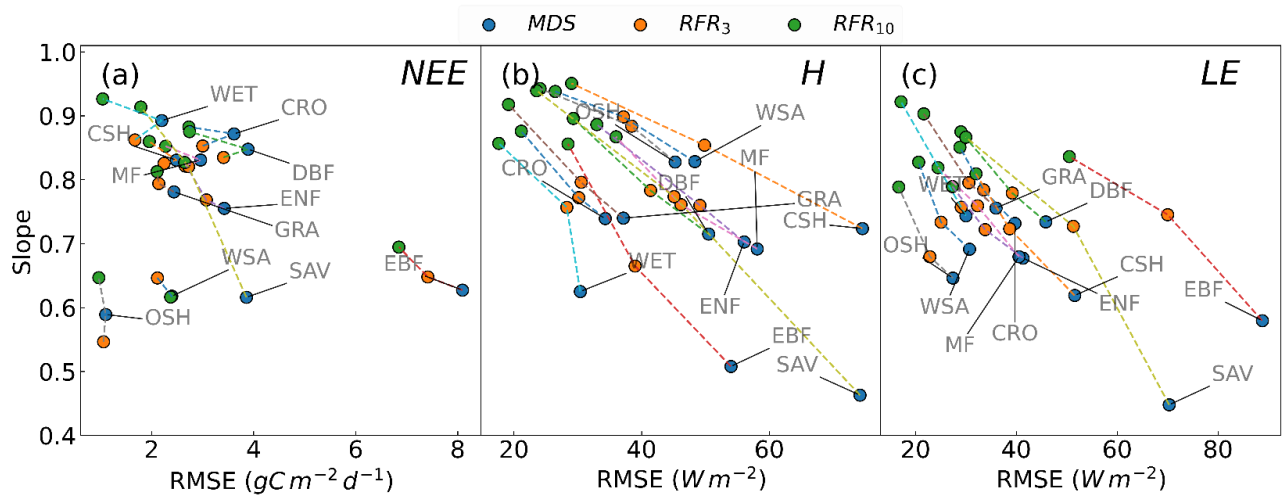


195 *Figure 4 R^2 and bias boxplots of MDS, RFR₃, and RFR₁₀ gap-filling for NEE (a, d), H (b, e), and LE (c, f), respectively. In this*
 196 *and following boxplots, bars show the third quartile, median and the first quartile as three bars on the boxes in descending*
 197 *order, while the black triangles indicate the mean.*

198
 199 RFR generally outperformed MDS gap filling for all fluxes with higher R^2 and narrower bias
 200 interquartile range (IQR) (Figure 4). RFR₃ was out performed by RFR₁₀ for R^2 but not for bias, where
 201 RFR₃ had a marginally lower bias for LE and NEE (Figure 4e and f).

202 Across all three fluxes (NEE, H, LE), median R^2 showed RFR₃ explaining 9 %, 16 %, and 18 %
 203 more variance than MDS, respectively, and the RFR₁₀ explaining a further 8 %, 16 %, and 13 %, respectively (Figure 4a-c and Table S3Error! Reference source not found.). More details can be
 204 found in Table S4.

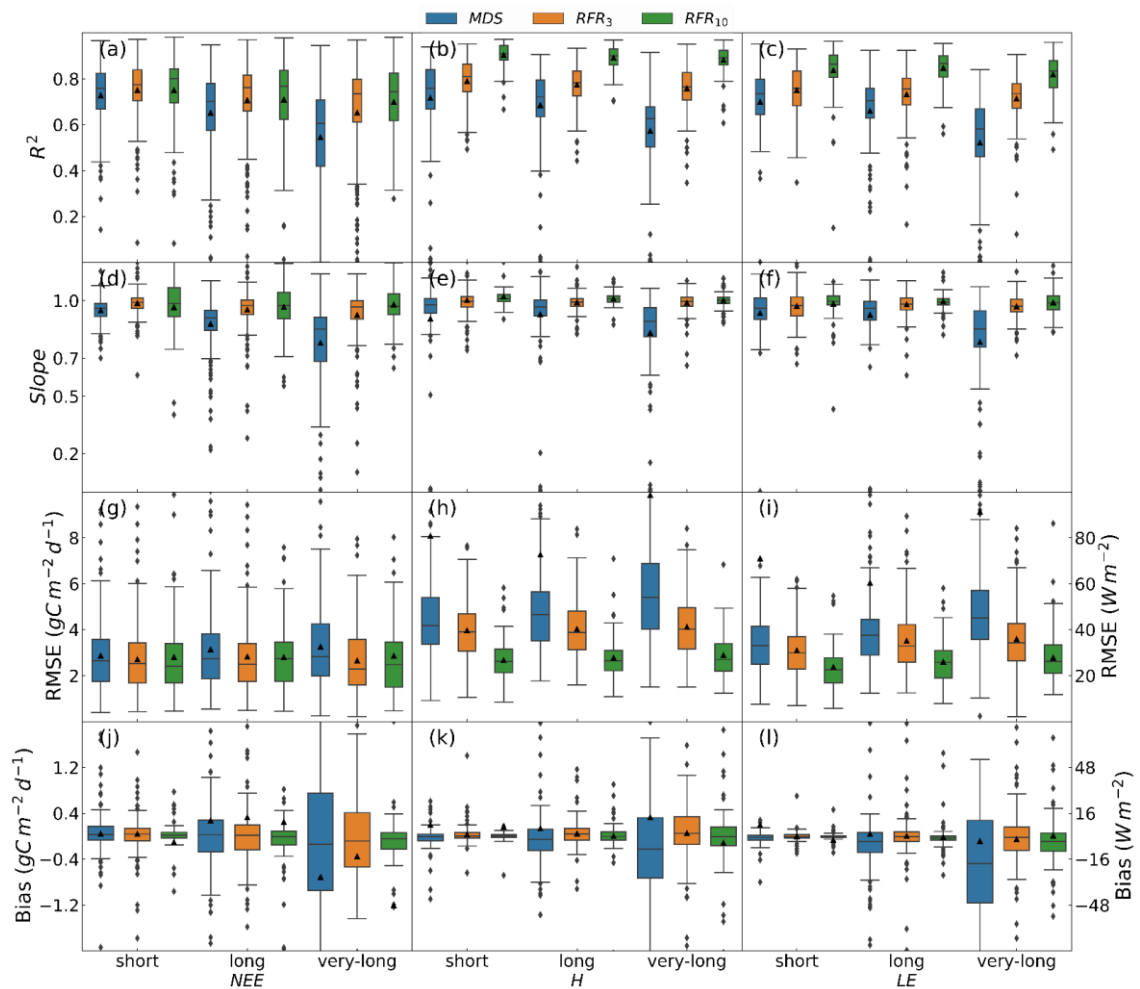
206 Both RFR₃ and RFR₁₀ resulted in similar reductions in the IQR of biases over MDS, nearly 40 %
 207 for NEE (Figure 4d) and more than 70 % for H and LE (Figure 4e and f). All methods showed a
 208 similar median bias (across all sites) for NEE, ranging from -0.02 to 0.01 g C m⁻² d⁻¹ (Table S3Error!
 209 Reference source not found.).



210 *Figure 5 Scatter plot of gap-filling RMSE against slope. Location of each dot represents the median of metrics for one gap-*
 211 *filling approach (blue for MDS, orange for RFR₃, and green for RFR₁₀) of one IGBP. Dots concentrating on the top-left corner*
 212 *reflect higher values of slope but smaller values of RMSE, vice versa. Dots for the same IGBP are collected by dashed lines*
 213 *(line colours differ by IGBP ecosystem classification). CRO: Croplands, CSH: Closed Shrublands, DBF: Deciduous Broadleaf*
 214 *Forests, EBF: Evergreen Broadleaf Forests, ENF: Evergreen Needleleaf Forests, GRA: Grasslands, MF: Mixed Forests, OSH:*
 215 *Open Shrublands, SAV: Savannas, WET: Permanent Wetlands, WSA: Woody Savannas.*

216
 217 Similar pattern of the gap-performance was seen, with RFR₁₀ performing better than RFR₃ and
 218 both RFRs performing better than MDS in terms of slope and RMSE for all three fluxes (NEE, LE
 219 and H) and all ecosystems (Figure 5). RFR₃ increased the slope by 5 % over MDS, with RFR₁₀ nearly
 220 doubling this to 11 %. Meanwhile, RFR₃ reduced the RMSE by 17 % compared to MDS and RFR₁₀
 221 reduced RMSE 21 % compared to MDS (Table S4).

222 The improvements in gap-filling slope and RMSE brought by RFR methods were larger for H
 223 (Figure 5b) and LE (Figure 5c) than for NEE (Figure 5a). The improvement of RFR₁₀ was particularly
 224 evident for H and LE in ecosystems that MDS (and even RFR₃) struggle with (e.g., SAV and EBF,
 225 Figure 5b-c). Compared to MDS, the slope for RFR methods increased 3 % for NEE, 16 % for H, and
 226 15 % for LE; corresponding RSME decreased 15 % for NEE, 34 % for H, and 26 % for LE (Table S4).

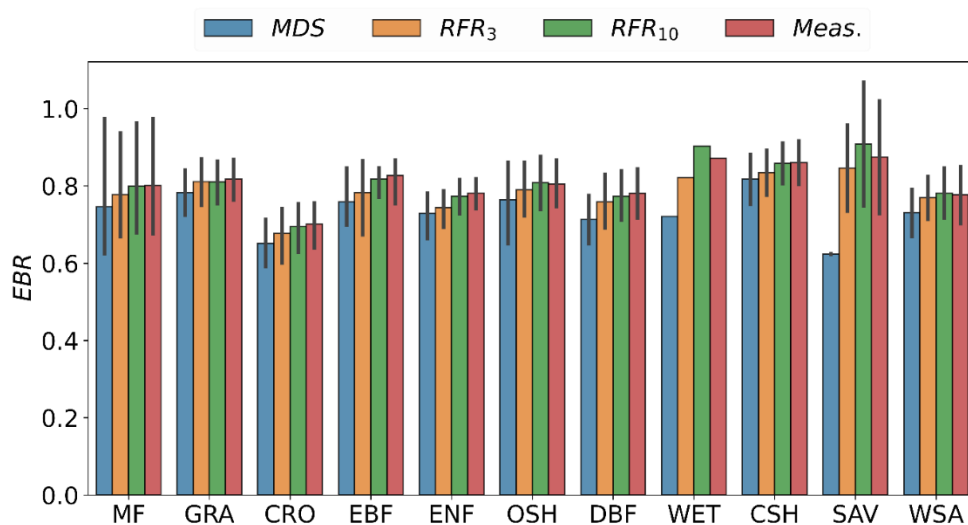


229 Figure 6 Boxplots showing gap-filling performance of three methods in short, long, and very-long gaps of same sites from the
 230 combined artificial gap scenario. The figures show the performance in terms of R^2 , slope, RMSE, and bias for NEE (a, d, g,
 231 and j), H (b, e, h, and k), and LE (c, f, i, and l), of R^2 (a - c), linear slope (d - f), RMSE (g - i), and bias (j - l).

232 Considering gap length scenarios separately, the RFR methods showed greater resilience to
 233 longer gaps compared to MDS (Figure 6). R^2 (Figure 6a-c) and slope (Figure 6d-f) of RFR₁₀ were
 234 higher than RFR₃ and further higher than MDS in short, long, and very-long gaps; while RMSE
 235 (Figure 6g-i) and IQR of bias (Figure 6j-l) of RFR₁₀ were smaller than RFR₃ and further smaller than
 236 MDS in short, long, and very-long gaps. More details can be found in Table S3 and S7.

237 All four statistical measures of the RFR methods were less sensitive to gap-length than MDS
 238 (Figure 6 and Table S3). For example, as gap length increased from short (1-day) to very-long (30-
 239 day), R^2 on average for the three fluxes decreased by 21 % (MDS), 4 % (RFR_3), and 4 % (RFR_{10}); gap-
 240 filling uncertainty in terms of bias interquartile range increased by 44 % (MDS), 42 % (RFR_3), and
 241 6 % (RFR_{10}). In addition, RFR methods for H and LE showed higher accuracy in filling longer gaps
 242 than for MDS (e.g., higher mean R^2 and narrower R^2 IQR, Figure 6a-c).

243

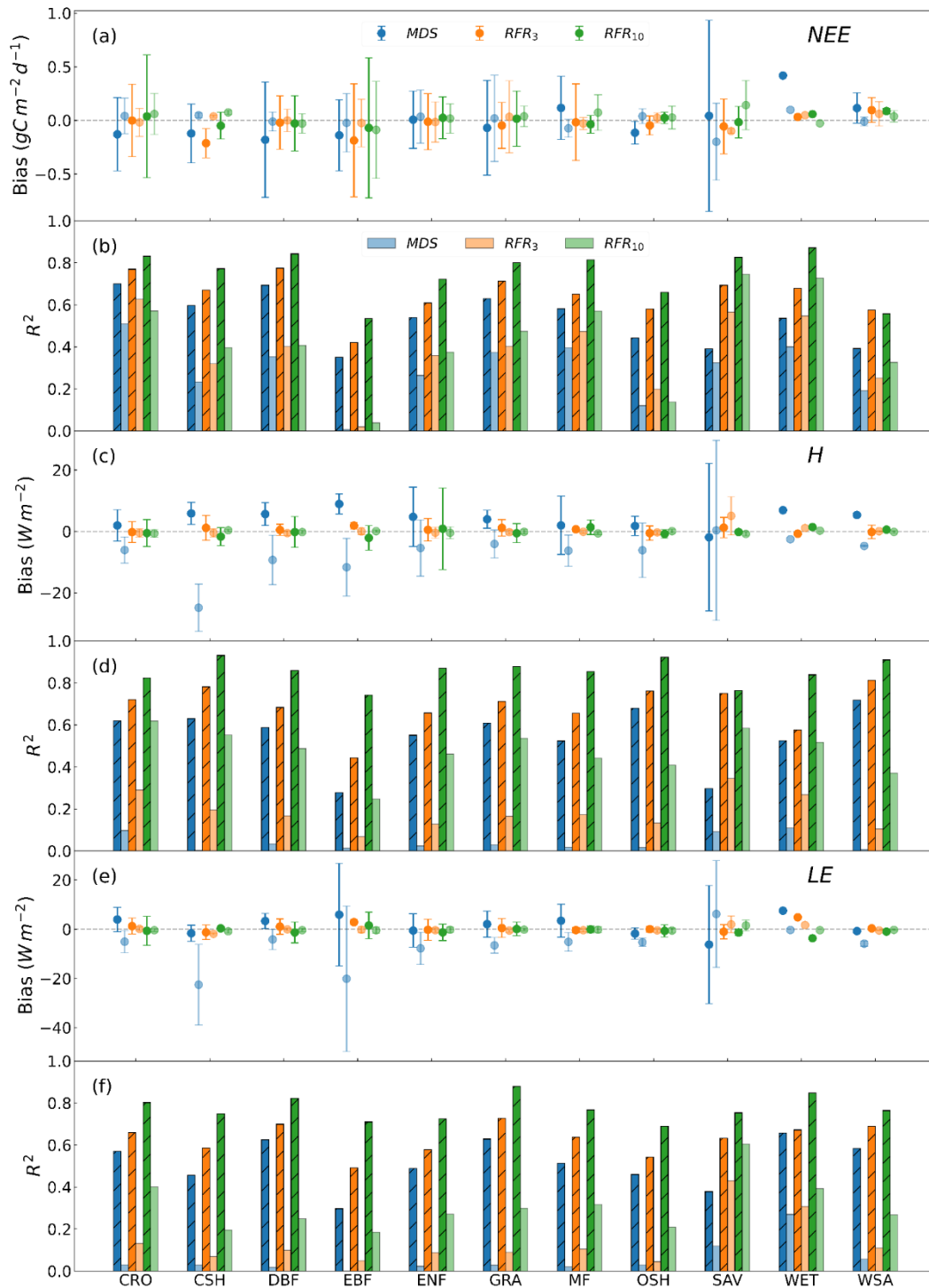


244 *Figure 7 Means (bars) and standard deviations (black vertical lines) of energy balance ratio (EBR) for filled artificial gaps*
 245 *and corresponding measurements (Meas.).*

246 Using filled artificial gaps (H and LE) and their measured counterparts, RFR methods (in
 247 particular RFR_{10}) exhibited energy balance ratios closer to those calculated for the corresponding
 248 original measurements than did MDS (Figure 7). The averaged EBR was separately 80 % (measured),
 249 80 % (RFR_{10}), 78 % (RFR_3), and 73 % (MDS). In regard to ecosystem types, overall EBR of croplands
 250 were smaller than other ecosystems. It was seen in all ecosystem types that RFR_{10} EBR was closer to
 251 measured values than RFR_3 and even closer to the measured values than MDS, such discrepancy in
 252 EBR between MDS and RFR methods was the largest in SAV (Figure 7).

253

254 3.3. Limitations of gap-filling approaches



255 *Figure 8 Day and night Gap-filling median bias (with error bars) and R^2 grouped by IGBP for NEE (a and b), H (c and d), and*
 256 *LE (e and f). The solid dots and bars are for daytime gap-filling, while the lighter dots and bars are for night-time gap-filling.*

257 Gap-filling performance, in terms of R^2 , in the daytime was much better than at night (Figure
 258 8). NEE, for example, median nighttime R^2 decreased compared to daytime by 80% (MDS), 70%
 259 (RFR₃) and 85% (RFR₁₀). It can be seen that the difference between daytime and night-time gap-
 260 filling R^2 for H (Figure 8d) and LE (Figure 8f) was larger than for NEE (Figure 8b). Bias in the daytime

261 NEE was larger than at night (Figure 8Error! Reference source not found.a), however no
262 consistent pattern was observed for H (Figure 8Error! Reference source not found.c) and LE
263 (Figure 8Error! Reference source not found.e). More details can be found in Table S5 and S6.

264 Performance of the gap-filling routines varied by IGBP ecosystem landcover classification.
265 Evergreen broadleaf forest (EBF) was seen with the lowest R^2 and large nocturnal bias for NEE
266 (Figure 8Error! Reference source not found.a and b), H (Figure 8Error! Reference source not
267 found.d), and LE (Figure 8Error! Reference source not found.e). Savannah (SAV) showed large
268 nocturnal biases for all three fluxes.

269

270 4. Discussion

271 4.1. Global gap-filling performance and intercomparison between approaches in different 272 landcover classifications

273 This work follows the earlier gap-filling study of NEE by Moffat et al. (2007), as well as H and
274 LE (Vitale et al. 2019), long gap uncertainty study (Richardson and Hollinger 2007), and recent
275 studies regarding high-performance of machine-learning on methane gap-filling (Kim et al. 2020;
276 Irvin et al. 2021). We updated and integrated previous analyses by applying machine-learning
277 approaches with modifications to fill very long gaps in NEE (CO_2), H, and LE fluxes and greatly
278 extended the geographical range of test sites. Our results showed a consistent improvement in
279 gap-filling using RFR compared to MDS for all the 94 global sites that were suitable for our
280 complete analysis (See methods). This improvement was seen for all three fluxes (NEE, LE and H),
281 with the greatest improvements for H and LE. For longer gaps, usually resulting from system failure
282 (Richardson and Hollinger 2007), the improvement on gap-filling by RFR could be considerable
283 (Figure 6 and Table S3Error! Reference source not found.), which supports the recommendation
284 for RFR given in Kim et al. (2020).

285 In agreement with previous studies, MDS showed satisfactory gap-filling performance in most
286 cases (Figure 6 and Table S3**Error! Reference source not found.**) because individual gap-lengths
287 are normally shorter than 1.5 days (Moffat et al. 2007). RFR methods improved the gap-filling
288 accuracy (e.g. 15 % R^2 increase by using RFR₃ and 30 % R^2 increase by using RFR₁₀) while reducing
289 uncertainties (e.g. interquartile range of bias decreased by 70 %) for NEE, H, and LE globally (Figure
290 3 and Table S3**Error! Reference source not found.**) and statistically significantly (Table S10) for
291 most of sites (Table S4-S6). Such improvement can be attributed to the complex architecture of
292 random forest and the “receptive-limiter” approach used in this study. The benefit of the
293 receptor-limiter we used can be seen by comparing random forest gap-filling performance with
294 and without the “receptive-limiter” (Figure S1).

295 The improvement for H and LE on gap-filling by using RFR was much larger than for NEE
296 compared with MDS (Figure 4). Currently, studies of gap-filling focused on H or LE are fewer than
297 NEE at a global scale (Foltýnová et al. 2020), resulting in a knowledge gap around these energy
298 fluxes. Reliable gap-filling methods (for H and LE) like RFR can help address this knowledge gap
299 and will help to inform debates around the environmental impacts (positive or negative) of nature-
300 based solutions and the mitigation of global climate change (Stenzel et al. 2018).

301 Using the extended driver set in RFR₁₀ showed advantages in gap-filling for R^2 , slope, and RMSE,
302 but the uncertainty also increased in some circumstances. Where the focus was solely on annual
303 sums – especially when only shorter gaps exist – RFR₃ produced the smallest range in biases. The
304 advantages of using extended drivers (RFR₁₀) became more apparent under the more challenging
305 gap scenarios (i.e. longer gaps and night-time).

306 Our analysis has shown a large variation in gap-filling performance for different ecosystems.
307 RFR indeed improved gap-filling performance, but it still struggled with NEE, H, and LE for
308 savannah (SAV), evergreen broadleaf forest (EBF), and open shrubland (OSH) (Figure 5) and
309 geographically in Africa, South America, and northwest North America (Figure 3). The reason

310 causing the poor gap-filling performance for ecosystems like EBF and ecosystems like SAV may be
311 different. Inferred by the small RMSE and slope, the poor performance in SAV could be accounted
312 by the weak flux signal there (Figure 5a). In contrast, the RMSE was large while the slope was small
313 for EBF (Figure 5a), which indicates the fluxes there could be large. The poor gap-filling
314 performance for EBF could be caused by the subtle seasonality, e.g. in Brazil, that does not
315 correlated with photosynthetically active radiation (Restrepo-Coupe et al. 2013). Given the large
316 improvement of using extended drivers, one possible solution in the future could be introducing
317 other environmental drivers, like leaf area index and/or satellite-based vegetation index, as
318 suggested by (Kang et al. 2019).

319 4.2. Gap-filling longer gaps and uncertainty analyses

320 The performance of MDS reduced significantly for very-long gaps, whereas RFR continued to
321 operate with similar statistical performance. Within our 94 selected sites (which are biased
322 towards complete datasets) MDS failed to gap-fill 5.47 % NEE half-hours from 19.50 % sites, 0.30 %
323 H half-hours from 13.07 % sites and 0.35 % LE half-hours from 13.73 sites. Crucially for NCS, RFR
324 did a better job at maintaining gap-filling performance for longer data gaps, for example, R^2 of
325 MDS in filling very-long gaps decreased by > 15 %, but the decrease for RFR methods was less than
326 5 % (Figure 6, **Error! Reference source not found.** Figure S2, and Table S3**Error! Reference source**
327 **not found.**).

328 Whilst both RFR methods outperformed MDS for long gaps, the performance of RFR₁₀ was
329 significantly better than RFR₃ (Figure 6). Where drivers are available RFR₁₀ should be considered
330 over RFR₃ or MDS for sites with data gaps that exceed a few days in length. It is worth noting
331 however that the average ratio of gap to data in the Fluxnet2015 (at the half hour resolution) is
332 67.53 % (i.e. on average datasets are missing 67.53% of their total half hours) and that of this
333 67.53%, 97.1% are short gaps, 2.77% are long gaps and 0.13% are very long gaps. Similarly, the
334 real gap ratio for H is 39.77 %, and 98.60 % are short gaps, only 1.20 % are long gaps and 0.20 %

335 are very-long gaps; the real gap ratio for LE is 44.99 %, and 98.87 % are short gaps, only 0.99 % are
336 long gaps and 0.14 % are very-long gaps. It might be suggested, however, that the data present in
337 FLUXNET are likely to represent 'best-case' data with contributions from better-maintained sites,
338 it is likely that gap scenarios may be more challenging at many other sites.

339 As an independent verification, the energy balance ratio (EBR) of 94 sites was 80 % (using
340 measured H and LE), 80 % (using RFR₁₀ gap-filled H and LE), 78 % (using RFR₃ gap-filled H and LE),
341 and 73 % (using MDS gap-filled H and LE); also suggesting the application of RFR methods can be
342 reliable in gap-filling energy fluxes. In this case, flux time-series gap-filled by using RFR methods
343 can be beneficial to climate models and/to support satellite remote sensing validations.

344 4.3. Implications of gap-filling performance for cumulative fluxes

345 In terms of gap-filling uncertainty, the mean global carbon sequestration rate is approximate
346 17.5 g C m⁻² yr⁻¹ for terrestrial ecosystems (Levin 2001; Griscom et al. 2017), and a week-long gap
347 would result in an additional uncertainty of 30 g C m⁻² yr⁻¹ in the worst cases (Richardson and
348 Hollinger 2007). Our findings suggest lower overall uncertainties, the bias interquartile range
349 across 94 sites equated to an annual bias of 84 g C m⁻² yr⁻¹ (MDS), 45 g C m⁻² yr⁻¹ (RFR₃), and 55 g
350 C m⁻² yr⁻¹ (RFR₁₀) (Table S3**Error! Reference source not found.**), that is comparable to Richardson
351 and Hollinger (2007). This reduction in NEE uncertainty by using RFR could be very valuable to
352 near carbon neutral ecosystems (Soloway et al. 2017). RFR methods also reduced uncertainty for
353 H and LE to <2 W m⁻² from 5 W m⁻² of MDS, and the improvement was good compared with > 3 W
354 m⁻² at most sites (Vitale et al. 2019). This reduction in uncertainty seen using RFR could play an
355 important role in accurately estimating global evapotranspiration. Therefore, RFR methods,
356 especially the RFR₃, are suggested with great potential in remote NCS applications where longer
357 gaps can occur more easily due to instrument failure. In remote areas, EC system maintenance in
358 a regular and frequent manner becomes difficult, as NCS applications aim to be low-cost.

359 4.4. Limitations of this study

360 RFR performed reliably in our study scenarios of gap lengths up to one month, but we might
361 expect performance to drop off substantially as gap lengths increase beyond this. We did not test
362 longer gaps due to the reduction in the numbers of FLUXNET sites that could be included in this
363 analysis but could usefully be the focus in a future study. Furthermore, as with other comparisons
364 studies such Moffat et al. (2007), we did not consider non-randomly located gaps in this study, for
365 example, gaps created due to regular maintenance schedules, or perhaps routine harvesting
366 operations in agricultural systems. Devising data gap probabilities based on potential
367 environmental and management challenges that were realistic across all 94 sites would be
368 extremely challenging. However, we suggest that focused studies looking at gap-filling
369 performance for non-random gaps could be an important focus for later studies.

370 The performance of gap-filling methods has been observed to be better during daytime than
371 night-time Moffat et al. (2007). Whilst our present study, RFR₁₀ performed slightly better than
372 RFR₃, and both improved on MDS, in capturing the diurnal patterns of NEE, the gap-filling
373 performance at night remains poor compared to daytime (e.g. $R^2 < 0.6$ in many ecosystems). One
374 reason is the low friction velocity at night, up-to 70 % of data can be rejected at night due to stable
375 atmospheric conditions etc.(Aubinet et al. 2012) and lower magnitude of nocturnal fluxes. In
376 addition, gap-filling at night is challenging because the shortwave solar radiation (vital to gap-
377 filling) vanishes (Reichstein et al. 2005).

378

379 5. Conclusion

380 In this study, a robust gap-filling approach (i.e. RFR) is proposed for filling long gaps in NEE, H,
381 and LE fluxes. Validated against MDS globally with gap sizes ranging from 1 to 30 days, we found
382 that RFR methods improve the gap-filling performance particularly for H and LE and extended
383 drivers are beneficial to gap-filling performance (i.e. RFR₁₀ outperforms RFR₃). RFR₃ and RFR₁₀
384 separately improves gap-filling accuracy by 15 % and 30 % while reduces uncertainty by 70 %.

385 Unlike MDS, RFR methods maintain performance with gap-lengths up to one month. Compared
386 with filling 1-day long gaps, the gap-filling performance (in terms of R^2) of filling 30-day long gaps
387 degrades by 21 % for MDS and degrades by < 4 % for RFR methods. No obvious difference is found
388 between RFR₃ and RFR₁₀ performance degradation. In addition, RFR methods, in particular the
389 RFR₁₀ largely reduces the uncertainty in filling 30-day long gaps, its uncertainty is less than 1/3 of
390 MDS. Three challenges are to be addressed in the future for better applying RFR gap-filling to eddy
391 covariance for natural climate solutions: 1) the difficulties of gap-filling at night which is a lasting
392 challenge to eddy covariance requires further research, 2) the still poor performance for certain
393 ecosystems (i.e. evergreen broadleaf forest, savannah, and open shrubland) that might be
394 addressed by introducing extra environmental drivers, 3) the question of gap-filling performance
395 for even longer gaps and non-random gaps that will be considered in our future studies.

396

397 6. Acknowledgements

398 The authors thank the FLUXNET and the research groups for providing the CC-BY-4.0 (Tier one)
399 open-access eddy covariance data (https://fluxnet.org/login/?redirect_to=/data/download-data/).
400 They also thank the ReddyProc (<https://cran.r-project.org/web/packages/REddyProc/index.html>)
401 team and scikit-learn (<https://scikit-learn.org/stable/install.html>) team for the packages that help the
402 implementation and validation for gap-filling approaches. S.Z. would like to acknowledge a Shell
403 funded PhD studentship and T.C.H. acknowledge funding from a joint UK NERC-FAPESP grant no.
404 NE/S000011/1 & FAPESP-19/07773-1.

405

406 7. Conflict of interests

407 We thank Shell for providing project funding to Tim Hill (University of Exeter) and a PhD
408 studentship grant award to Songyan Zhu (University of Exeter).

409

410 8. Reference

411 Aubinet M, Vesala T, Papale D (2012) Eddy covariance: a practical guide to measurement and data
412 analysis. Springer Science & Business Media

413 Baldocchi D (2014) Measuring fluxes of trace gases and energy between ecosystems and the
414 atmosphere--the state and future of the eddy covariance method. *Glob Change Biol*
415 20:3600–3609

416 Baldocchi D, Falge E, Gu L, et al (2001) FLUXNET: A new tool to study the temporal and spatial
417 variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. *Bull*
418 *Am Meteorol Soc* 82:2415–2434

419 Baldocchi DD (2020) How eddy covariance flux measurements have contributed to our
420 understanding of global change biology. *Glob Change Biol* 26:242–260

421 Bautista N, Marino BD, Munger JW (2021) Science to Commerce: A Commercial-Scale Protocol for
422 Carbon Trading Applied to a 28-Year Record of Forest Carbon Monitoring at the Harvard
423 Forest. *Land* 10:163

424 Breiman L (2001) Random forests. *Mach Learn* 45:5–32

425 Campioli M, Malhi Y, Vicca S, et al (2016) Evaluating the convergence between eddy-covariance and
426 biometric methods for assessing carbon budgets of forests. *Nat Commun* 7:13717.
427 <https://doi.org/10.1038/ncomms13717>

428 Chausson A, Turner B, Seddon D, et al (2020) Mapping the effectiveness of nature-based solutions
429 for climate change adaptation. *Glob Change Biol* 26:6134–6155.
430 <https://doi.org/10.1111/gcb.15310>

431 Cohen-Shacham E, Andrade A, Dalton J, et al (2019) Core principles for successfully implementing
432 and upscaling Nature-based Solutions. *Environ Sci Policy* 98:20–29.
433 <https://doi.org/10.1016/j.envsci.2019.04.014>

434 Derrick B, Toher D, White P (2016) Why Welch's test is Type I error robust. *Quant Methods Psychol*
435 12:

436 Eshonkulov R, Poyda A, Ingwersen J, et al (2019) Evaluating multi-year, multi-site data on the energy
437 balance closure of eddy-covariance flux measurements at cropland sites in southwestern
438 Germany. *Biogeosciences* 16:521–540. <https://doi.org/10.5194/bg-16-521-2019>

439 Falge E, Baldocchi D, Olson R, et al (2001) Gap filling strategies for defensible annual sums of net
440 ecosystem exchange. *Agric For Meteorol* 107:43–69

441 Foken T, Aubinet M, Finnigan JJ, et al (2011) Results of a panel discussion about the energy balance
442 closure correction for trace gases. *Bull Am Meteorol Soc* 92:13–18.
443 <https://doi.org/10.1175/2011BAMS3130.1>

444 Foltýnová L, Fischer M, McGloin RP (2020) Recommendations for gap-filling eddy covariance latent
445 heat flux measurements using marginal distribution sampling. *Theor Appl Climatol* 139:677–
446 688. <https://doi.org/10.1007/s00704-019-02975-w>

447 Friedlingstein P, O'Sullivan M, Jones MW, et al (2020) Global Carbon Budget 2020. *Earth Syst Sci Data*
448 12:3269–3340. <https://doi.org/10.5194/essd-12-3269-2020>

449 Glanemann N, Willner SN, Levermann A (2020) Paris Climate Agreement passes the cost-benefit test.
450 *Nat Commun* 11:1–11. <https://doi.org/10.1038/s41467-019-13961-1>

451 Griscom BW, Adams J, Ellis PW, et al (2017) Natural climate solutions. *Proc Natl Acad Sci* 114:11645–
452 11650

453 Hemes KS, Runkle BRK, Novick KA, et al (2021) An Ecosystem-Scale Flux Measurement Strategy to
454 Assess Natural Climate Solutions. *Environ Sci Technol* 55:3494–3504.
455 <https://doi.org/10.1021/acs.est.0c06421>

456 Hill T, Chocholek M, Clement R (2017) The case for increasing the statistical power of eddy
457 covariance ecosystem studies: why, where and how? *Glob Change Biol* 23:2154–2165

458 Irvin J, Zhou S, McNicol G, et al (2021) Gap-filling eddy covariance methane fluxes: Comparison of
459 machine learning model predictions and uncertainties at FLUXNET-CH4 wetlands. *Agric For*
460 *Meteorol* 308–309:108528. <https://doi.org/10.1016/j.agrformet.2021.108528>

461 Jung M, Reichstein M, Bondeau A (2009) Towards global empirical upscaling of FLUXNET eddy
462 covariance observations: Validation of a model tree ensemble approach using a biosphere
463 model. *Biogeosciences* 6:2001–2013

464 Kang M, Ichii K, Kim J, et al (2019) New Gap-Filling Strategies for Long-Period Flux Data Gaps Using a
465 Data-Driven Approach. *Atmosphere* 10:568

466 Keith H, Vardon M, Obst C, et al (2021) Evaluating nature-based solutions for climate mitigation and
467 conservation requires comprehensive carbon accounting. *Sci Total Environ* 769:144341–
468 144341. <https://doi.org/10.1016/j.scitotenv.2020.144341>

469 Kim Y, Johnson MS, Knox SH, et al (2020) Gap-filling approaches for eddy covariance methane fluxes:
470 A comparison of three machine learning algorithms and a traditional method with principal
471 component analysis. *Glob Change Biol* 26:1499–1518

472 Levin SA (2001) *Encyclopedia of biodiversity*

473 Mauder M, Foken T, Clement R, et al (2008) Quality control of CarboEurope flux data--Part 2: Inter-
474 comparison of eddy-covariance software. *Biogeosciences* 5:451–462

475 Menzer O, Meiring W, Kyriakidis PC, McFadden JP (2015) Annual sums of carbon dioxide exchange
476 over a heterogeneous urban landscape through machine learning based gap-filling. *Atmos*
477 *Environ* 101:312–327. <https://doi.org/10.1016/j.atmosenv.2014.11.006>

478 Moffat AM, Papale D, Reichstein M, et al (2007) Comprehensive comparison of gap-filling techniques
479 for eddy covariance net carbon fluxes. *Agric For Meteorol* 147:209–232

480 Papale D, Reichstein M, Aubinet M, et al (2006) Towards a standardized processing of Net Ecosystem
481 Exchange measured with eddy covariance technique: algorithms and uncertainty estimation.
482 *Biogeosciences* 3:571–583

483 Pastorello G, Papale D, Chu H, et al (2017) A new data set to keep a sharper eye on land-air
484 exchanges. *Eos Trans Am Geophys Union Online* 98:

485 Pastorello G, Trotta C, Canfora E, et al (2020) The FLUXNET2015 dataset and the ONEFlux processing
486 pipeline for eddy covariance data. *Sci Data* 7:1–27

487 Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: Machine learning in Python. *J Mach*
488 *Learn Res* 12:2825–2830

489 Perez-Priego O, El-Madany TS, Migliavaca M, et al (2017) Evaluation of eddy covariance latent heat
490 fluxes with independent lysimeter and sapflow estimates in a Mediterranean savannah
491 ecosystem. *Agric For Meteorol* 236:87–99. <https://doi.org/10.1016/j.agrformet.2017.01.009>

492 Reichstein M, Falge E, Baldocchi D, et al (2005) On the separation of net ecosystem exchange into
493 assimilation and ecosystem respiration: review and improved algorithm. *Glob Change Biol*
494 11:1424–1439

495 Restrepo-Coupe N, Da Rocha HR, Hutrya LR, et al (2013) What drives the seasonality of
496 photosynthesis across the Amazon basin? A cross-site analysis of eddy flux tower
497 measurements from the Brasil flux network. *Agric For Meteorol* 182:128–144

498 Richardson AD, Hollinger DY (2007) A method to estimate the additional uncertainty in gap-filled
499 NEE resulting from long gaps in the CO₂ flux record. *Agric For Meteorol* 147:199–208.
500 <https://doi.org/10.1016/j.agrformet.2007.06.004>

501 Skinner RH, Dell CJ (2015) Comparing pasture C sequestration estimates from eddy covariance and
502 soil cores. *Agric Ecosyst Environ* 199:52–57

503 Smith P, Beaumont L, Bernacchi CJ, et al (2021) Essential outcomes for COP26. *Glob Change Biol*

504 Smith P, Soussana J-F, Angers D, et al (2020) How to measure, report and verify soil carbon change
505 to realize the potential of soil carbon sequestration for atmospheric greenhouse gas
506 removal. *Glob Change Biol* 26:219–241

507 Soloway AD, Amiro BD, Dunn AL, Wofsy SC (2017) Carbon neutral or a sink? Uncertainty caused by
508 gap-filling long-term flux measurements for an old-growth boreal black spruce forest. *Agric*
509 *For Meteorol* 233:110–121. <https://doi.org/10.1016/j.agrformet.2016.11.005>

510 Stenzel F, Greve P, Tramberend S (2018) increase water stress more than climate change. *Nat*
511 *Commun* 1–9. <https://doi.org/10.1038/s41467-021-21640-3>

512 Tramontana G, Ichii K, Camps-Valls G, et al (2015) Uncertainty analysis of gross primary production
513 upscaling using Random Forests, remote sensing and eddy covariance data. *Remote Sens*
514 *Environ* 168:360–373

515 Vitale D, Bilancia M, Papale D (2019) A multiple imputation strategy for eddy covariance data. *J Env*
516 *Inf* 34:68–87

517 Wollenberg E, Richards M, Smith P, et al (2016) Reducing emissions from agriculture to meet the 2 °C
518 target. *Glob Change Biol* 22:3859–3864. <https://doi.org/10.1111/gcb.13340>

519 Wutzler T, Lucas-Moffat A, Migliavacca M, et al (2018) Basic and extensible post-processing of eddy
520 covariance flux data with REddyProc. *Biogeosciences* 15:5015–5030

521 Zeng J, Matsunaga T, Tan Z-H, et al (2020) Global terrestrial carbon fluxes of 1999–2019 estimated
522 by upscaling eddy covariance data with a random forest. *Sci Data* 7:1–11

523

524

525