Environmental observatory networks (EONs) provide information to understand and forecast the spatial and temporal dynamics of Earth’s biophysical processes. Consequently, representativeness analyses are important to provide insights for improving EONs’ management, design, and interpretation of their value-added products. We assessed the representativeness of registered FLUXNET sites (n = 41, revised on September 2018) across Latin America (LA), a region of great importance for the global carbon and water cycles, which represents 13% of the world’s land surface. Nearly 46% of registered FLUXNET sites are located in evergreen broad-leaf forests followed by sites in woody savannas (~20%). Representativeness analyses were performed using a 0.05° spatial grid for multiple environmental variables, gross primary productivity (GPP), and evapotranspiration (ET). Our results showed a potential representativeness of 34% of the surface area for climate properties, 36% for terrain parameters, 34% for soil resources, and 45% when all aforementioned environmental variables were summarized into a principal component analysis. Furthermore, there was a 48% potential representativeness for GPP and 34% for ET. Unfortunately, data from these 41 sites are not all readily available for the scientific community, limiting synthesis studies and model benchmarking/parametrization. The implication is that global/regional data-driven products are forced to use information from FLUXNET sites outside LA to predict patterns in LA. Representativeness could increase to 86% (for GPP) and 80% (for ET) if 200 sites are optimally distributed. We discussed ongoing challenges, the need to enhance interoperability and data sharing, and promote monitoring efforts across LA to increase the accuracy of regional-to-global data-driven products.
designed to provide insights to address complex regional-to-global socio-ecological problems through a coordinated effort (Chabbi et al., 2017; Keller et al., 2011; Scholes et al., 2017). Some key tasks lead by EONs include data collection, data sharing and synthesis activities that are useful for scientific discovery and making informed environmental policy or management decisions (Lovett et al., 2007; Scholes et al., 2017; Villarreal et al., 2018).

An example of an EON is FLUXNET, which represents a global network of study sites using the eddy-covariance method to measure the exchange of mass and energy between the land surface and the atmosphere (D. D. Baldocchi, 2020; D. Baldocchi et al., 2001). FLUXNET is a global “network of regional networks” that promotes compilation, harmonization, standardization, archiving, and synthesis activities of eddy-covariance data. The FLUXNET network is present across multiple ecosystems around the world, so it is possible to generate knowledge of the interaction between terrestrial ecosystems and the atmosphere from regions to the global scale (Falge et al., 2002; I. B.; J. B. Fisher et al., 2008; Keenan et al., 2014; Schwalm et al., 2017). However, FLUXNET sites are neither randomly nor systematically distributed, so they underrepresent certain regions and ecosystems across the world (D. D. Baldocchi, 2020; Papale et al., 2015; Villarreal et al., 2019). Consequently, representativeness assessments of EONs are critical as they provide information for EONs design/growth, and insights for interpretations and implications of data-driven (or value-added) products (Sulkava et al., 2011; Villarreal et al., 2018). These assessments are relevant to increase EONs’ applicability and to guide regional-to-global management and research efforts (Jongman et al., 2017; Lovett et al., 2007).

The representativeness of EONs has been mostly assessed using climate and vegetation parameters (Hargrove et al., 2003; Sulkava et al., 2011). For example, through stratification of climate, vegetation, and soil information the representativeness of AmeriFlux has been assessed (Hargrove et al., 2003), while recent studies have incorporated functional information from ecosystems (Alcaraz-Segura et al., 2017; Villarreal et al., 2018, 2019). A common approach to assess the representativeness of EONs has been the estimation of minimum distances within a multivariate space (Hargrove et al., 2003; Sulkava et al., 2011). An alternative approach is the use of machine learning techniques, which estimate the spatial distribution of the environmental range monitored by the EON’s study sites (i.e., nodes) across the spatial domain of the network (Villarreal et al., 2018, 2019).

We propose that it is possible to assess the representativeness of EONs based on concepts derived from species distributions models (SDMs). Briefly, SDMs define a geographic space that includes a set of environmental data layers, and then delineate an area within the geographic space that corresponds to environmental properties that are suitable for the presence of a certain species (Evans et al., 2011). We propose that this concept can be applied to assess the representativeness of EONs since the goal is to delineate the spatial distribution of environmental factors across a geographic space that should be similar to the environmental range monitored by corresponding monitoring sites within an EON (Villarreal et al., 2018).

Here, we present a representativeness assessment of eddy-covariance sites registered with FLUXNET across Latin America (LA). LA is a region that is largely characterized by its wide ecosystem diversity along with a broad gradient of land-use and land-use-change types; especially when compared with other regions of the world (e.g., Europe or North America). Furthermore, LA is an important region for the global carbon and water cycles as it includes the Amazon (Briene et al., 2015) and important mountain ranges that contribute to the “water towers” of the world (Immerzeel et al., 2020). LA includes nearly 13% of the global land surface area but only about 5% of all registered FLUXNET sites are located within this region (estimated for year the 2018 but see methods for details). The density of registered FLUXNET sites in LA is very low when compared to regions such as the United States or Europe, and global data-driven products (e.g., FluxCom) are forced to use information from FLUXNET sites outside LA to predict patterns in LA. Hence, a representativeness analysis is needed to better interpret the available information within LA and the output of regional-to-global data-driven products parameterized with FLUXNET data.

The overarching goal of this study is to provide an assessment of the representativeness of registered FLUXNET sites across LA to monitor environmental factors such as climate, topography and soil resources along with ecosystem processes such as gross primary productivity (GPP) and evapotranspiration (ET). We asked four interrelated research questions: (1) What is the representativeness of FLUXNET sites across LA to characterize the spatial variability of climate, topography, and soil resources? (2) What is the representativeness of FLUXNET...
sites to monitor GPP and ET patterns across LA? (3) How does the representativeness of FLUXNET sites (to monitor GPP and ET) varies as spatial scale changes? and (4) How many more sites are needed to substantially improve the representativeness of GPP and ET across this region? Finally, this study is based on publicly available information and open-source software, so this framework can be applied anywhere across the world.

2. Data and Methods

2.1. FLUXNET Registered Sites

FLUXNET provides standardized data products through coordination among multiple regional eddy covariance networks across the globe (http://fluxnet.fluxdata.org). We used this online database to extract the geographical location of eddy-covariance sites across LA registered with FLUXNET. We identified 41 registered sites distributed across different ecosystems (revised on September 2018; Figure 1, Table S1) and we considered these sites for further analyzes despite they are active or inactive and if they have provided data or not to the FLUXNET database. Consequently, this study should be considered as a best-case scenario and provides a potential representation of eddy-covariance sites across LA. We hope that this study will encourage principal investigators to register their sites and share data with FLUXNET to improve the representation of LA in regional and global studies.

We recognize that there are several challenges and assumptions for performing an accurate representativeness assessment of eddy-covariance sites across LA. First, the assumption of 41 registered FLUXNET sites for the

![Figure 1](image-url). Spatial distribution of different land classes across Latin America according to the International Geosphere-Biosphere Program (IGBP), and the location of eddy-covariance sites affiliated with FLUXNET. See supplementary material (Table S2) for details about FLUXNET sites. Note that some sites overlap at this spatial scale due to their close proximity.
2018 does not mean that those sites are active nor that their data is or will be available for the scientific community. Second, there are unregistered eddy-covariance sites across LA and new sites are being installed or have become inactive. We emphasize that our assessment is a potential representation because data from these 41 sites are not readily available for the scientific community. For example, the FLUXNET2015 data set only includes 7 eddy covariance sites across LA (i.e., BR-Sa1, BR-Sa2, GF-Guy, PA-SPn, PA-SPs, AR-SLu, and AR-Vir), and the AmeriFlux network (revised August 2020) has 23 registered sites but only 8 of them share data with the network (Table S2). We did not include Caribbean islands for our assessment due to the relatively small spatial extent of this region (compared to the continental area of LA) and because no sites were registered with AmeriFlux or FLUXNET in 2018 to perform a formal analysis across this region.

### 2.2. Environmental Factors

A set of variables related to climate, terrain parameters, and soil resources variability were used to assess the representativeness of environmental state factors, as they constrain the spatial patterns of ecosystem processes such as GPP and ET (Amundson, 1991; Chapin et al., 2002). We used 19 bioclimatic predictors to characterize climate conditions: mean annual conditions (i.e., annual mean temperature, annual precipitation), mean annual seasonal conditions (i.e., temperature seasonality), and intraannual seasonal conditions (i.e., mean temperature of the driest quarter or precipitation of the wettest quarter) of temperature and precipitation (Hijmans et al., 2005). Terrain parameters were characterized by slope, elevation, topographic wetness index (used to quantify topographic influence on hydrology processes [Sørensen et al., 2006]), and solar radiation index. Soil resources were characterized by soil organic carbon, soil nitrogen, soil phosphorus, and soil water content. The bioclimatic predictors were downloaded from worldclim.org (accessed May 2018). Most terrain parameters and soil resources variables were downloaded from worldgrids.org (accessed May 2018), but soil organic carbon was downloaded from www.fao.org (accessed May 2018) and soil phosphorus from data.nasa.gov (accessed May 2018). Respectively, NASA-MODIS products MOD17A2 and MOD16A2 from 2001 to 2014 were used to characterize GPP and ET as previously done for assessment.
of the AmeriFlux and MexFlux networks (Villarreal et al., 2018, 2019). The statistic parameters used to characterize GPP and ET dynamics were the mean (GPP_mean, ET_mean) and the coefficient of variation (GPP_CV, ET_CV), since they have been used as proxies to represent ecosystem productivity and seasonality, respectively (Alcaraz-Segura et al., 2017; Villarreal et al., 2018, 2019).

2.3. Data Harmonization and FLUXNET Representativeness

All variables were standardized into a similar geographical system (GS), which consisted in harmonizing all variables into the same projection (i.e., WGS84) and transforming them into the same spatial resolution (i.e., 0.05°). We selected 0.05° as this resolution is largely used to represent environmental patterns at a regional scale (Chrysoulakis et al., 2003; Löw et al., 2011) and has been used to assess the representativeness of AmeriFlux, MexFlux and the National Ecological Observatory Network (NEON; Villarreal et al., 2018, 2019). In addition, all variables representing climate, terrain parameters and soil resources were reduced in dimensionality using a principal component analysis (PCA) to assess the representativeness of these combined environmental factors (using the first two principal components) from a multivariate approach. Representativeness was performed for all environmental parameters at 0.05°, while GPP and ET representativeness were estimated at 0.05°, 0.25°, 0.50°, and 1.0°, since global models of GPP and ET are usually estimated based on these spatial resolutions (R. A. Fisher & Koven, 2020).

Representativeness was estimated using random forest (RF) applied for SDMs. RF is a widely used technique in SDMs, especially for rare species that have few observations over a broad region (Cutler et al., 2007; Evans et al., 2011). We propose that the relative few numbers of eddy-covariance sites across the large geographic extent of LA is a similar case study. As a machine-learning technique, RF produces classification trees from bootstrapping samples from a given data set (i.e., training-data), while the observations that are not considered (out-of-bag data) are later used for predictions and model evaluation. First, classifications trees (CTs) are built from sample bootstrapping by repeatedly partitioning the training-data into a binary-series of clusters (i.e., child-nodes) that split the data into more or less homogeneous child-nodes with respect to the response variable, this process continues with each child-node until stops (Marmion et al., 2009). Second, the grown trees are used to predict the out-of-bag observations. The class that is predicted of observation is the majority vote of the out-of-bag predictions for that observation (Cutler et al., 2007; Evans et al., 2011; Marmion et al., 2009). Finally, RF produces a raster map that represents the relative similarity of each pixel to the sample points or presence data (Schmitt et al., 2017), which in this case corresponds to the geographic locations of FLUXNET sites across LA.

Model performance was assessed using True Skin Statistics (TSS), which corresponds to the sum of the model sensitivity (i.e., the proportion of presence correctly predicted) and the specificity (i.e., the proportion of absence correctly predicted) minus one. TSS ranges from −1 to 1, being −1 a predictability power worse than random model, 0 indicates a random predictability, and 1 corresponds to a perfect model (Liu et al., 2011). Absence data points were generated by random selection, as randomly selected points usually produce reliable distribution models (Barbet-Massin et al., 2012). Also, the uncertainty of each model was assessed by repeating 10 times each model with only one iteration and calculate their mean and standard deviation, as it was performed in a previous study (Villarreal et al., 2019). We assessed the 95% confidence interval to determine differences between represented and nonrepresented areas for the different environmental variables.

The optimal number of absence data and model repetition for each environmental set of variables (i.e., climate, terrain parameters, soil resources, GPP, and ET) was selected based on the TSS by an iterative process. We selected the number of absence data and model repetition that had higher TSS (Barbet-Massin et al., 2012). The number of absence and repetitions were different for each environmental set of variables.

Data management and analysis were performed using the R programming language (R project for statistical computing; http://www.r-project.org) and the ‘SSDM’ library (Schmitt et al., 2017).

2.4. Improving the Representativeness of GPP and ET Across LA

A final goal was to provide insights about how many more sites are needed across LA to improve the representativeness of GPP and ET. To this end, we used the constrained Latin hypercube sampling technique (cLHS;
Minasny & McBratney, 2006). The cLHS is a multivariate statistical technique that ensures full coverage of the range of the variables involved in the multivariate space. For this study, we used the mean and standard deviation of GPP and ET as discussed earlier. The cLHS serves as an efficient sampling strategy and it has been previously used for EON’s representativeness analysis (Villarreal et al., 2019). For this assessment we followed a sequential approach: (a) we started by adding additional sites across LA in increments of 10 sites until reaching 100 sites; then (b), additional sites were added in increments of 20 until reaching 200 sites across LA. We stopped at 200 potential sites across LA as an arbitrary number equivalent to the approximate total number of eddy-covariance sites registered in the AmeriFlux network for the conterminous United States. The assessment of representativeness by adding new sites was also performed using RF as described above.

3. Results

3.1. Distribution of FLUXNET Sites Across LA

There is a large diversity of terrestrial ecosystems across LA with 15 out of 17 possible International Geosphere-Biosphere Program (IGBP) categories (MODIS MCD12Q1, 2012). These IGBP categories are not evenly distributed neither the number of FLUXNET sites within each category (Table 1). For example, 5 out of 17 categories incorporate 80% of LA’s land surface, where the largest categories are Evergreen Broadleaf Forest (34% of LA) and Savanna (19% of LA). The categories with the least spatial extent are Evergreen Needle-Leaf Forest (0.05% of LA) and Closed Shrublands (0.04% of LA). The categories with the largest number of registered FLUXNET sites are Evergreen Broadleaf Forest (n = 19) followed by Savanna (n = 8), but we highlight that eight IGBP categories do not have any registered sites (Table 1).

3.2. Representativeness of Environmental Factors, GPP, and ET

The representativeness of FLUXNET sites (using a spatial grid of 0.05°) differed for each environmental factor (i.e., climate, terrain properties, soil resources, combined environmental properties, GPP, and ET) and between IGBP categories. The highest spatial representativeness corresponded to GPP (48%), combined environmental factors (45%; derived from the PCA), and terrain parameters (36%); while climate, soil resources, and ET had similar representativeness (34%; Table S3). The highest representativeness among IGBP categories corresponded to Shrublands and Savannas, while forest ecosystems (Evergreen and Deciduous Broadleaf Forest) had similar values as those from managed ecosystems (e.g., Croplands; Table S4).

From the representativeness models, we identified the two most important variables for each environmental factor and assessed the differences between represented and nonrepresented regions (Table S4). For bioclimatic predictors (Figures 2a and 2b), precipitation seasonality above 120 mm and below 40 mm and the annual mean diurnal temperature range above 20°C and below 6°C were not represented (Figures 2c and 2d). For terrain properties, the majority of IGBP classes were represented, while values 90<TWI<75 were not represented for TWI (topographic wetness index; Figures 2e and 2f). For soil resources (Figures 2g and 2h), soil organic carbon >80 g/m² and soil nitrogen below <500 and above >2,000 mg/m² were not represented (Figures 2i and 2j). For all the environmental drivers combined into a PCA, the variables that had the highest influence were terrain complexity (21% of total variability) and soil nitrogen (6% of total variability). Overall, PC1 was represented within the range ~0.08 to 0.16 (Figures 2k and 2l) while PC2 was represented within the range of ~1 to 1.5 (Figures 2m and 2n). Detailed statistics for represented and nonrepresented areas are in Table S5.

The mean for GPP_mean and GPP_CV across LA were 4.3 and 1.24 g C/m² day, respectively. The representativeness of FLUXNET sites for GPP_mean was biased toward values >4g C/m² day, while for GPP_CV values > 2 g C/m² day were not represented. Nonrepresented areas had mean values of GPP_mean and GPP_CV of 4.21 and 1.16 g C/m² day, respectively (Table S5; Figures 3a–3d).

The mean values for ET_mean and ET_CV across LA were 2.51 and 0.71 mm/day, respectively. The ET_mean and ET_CV monitored by FLUXNET sites were 2.42 and 0.67 mm/day, respectively (Figures 3e–3f). Nonrepresented areas had mean values of ET_mean and ET_CV of 2.42 and 0.68 mm/day, respectively (Figures 3g–3h).

We present the spatial representativeness for the two most important variables for each environmental factor (Figure 4). Climate-related variables are not represented across most of the southern part of LA (Figure 4a), although registered FLUXNET sites cover a wide range of precipitation and temperature across LA.
climate space (Figure 4b). The representation of terrain parameters and soil resources are scattered across LA (Figures 4c and 4e). Although the range of TWI is generally covered (Figure 4d), registered FLUXNET sites seem to be concentrated in a narrow range of areas with similar soil nitrogen and soil carbon properties (Figure 4f). Finally, when all environmental variables were assessed using a PCA it is evident that most sites are clustered within similar environmental characteristics (Figure 4i) with lack of representation of the southern part of LA, the Caatinga in Brazil, and the northern arid lands of Mexico (Figure 4g).

The spatial representation of GPP (using a spatial grid of 0.05°) is 48% but it is scattered across LA. We highlight that the Caatinga, southern parts of the Cerrado and Atlantic Forest in Brazil are not represented (Figure 5a). Similarly, the Pantanal and Pampa are regions that are not represented for GPP. We identified that there is a cluster of registered FLUXNET sites within relatively large GPP_mean (~7.5 g C/m² day) and a GPP_CV of ~1.5 g C/m² day (Figure 5b).

The spatial representation of ET (using a spatial grid of 0.05°) is only 34% and more scattered across LA. Large areas of the northern Amazon Forest, the Caatinga, Cerrado, Pantanal and the Pampa are not represented by the registered FLUXNE sites (Figure 5c). Although sites are more scattered for representing ET dynamics, we still found a cluster of registered FLUXNET sites with a relatively high ET_mean (~4 mm/day) but lower ET_CV ~0.6 mm/day; Figure 5d).

The representativeness of GPP and ET at 0.25°, 0.50°, and 1.0° slightly decreased as the spatial resolution became coarser. For example, GPP representativeness was 24%, 21% and 18% while ET had 22%, 19% and 16% at 0.25°, 0.50°, and 1.0°, respectively. The decrease in representativeness was partially associated to a decrease in the number of effective registered FLUXNET sites (35, 31, and 30 effective sites) across LA. The reason is that as the spatial resolution decreases (i.e., pixels are larger) some sites are included within the larger pixels due to the spatial proximity among them.

3.3. Representativeness by Adding Potential New Sites and Uncertainty Assessment

Theoretically, the overall representativeness across LA increased from 45% to 86% for GPP and from 42% to 80% for ET by adding up to 200 study sites (Figures 6a and 6b; Table 2). These sites should be optimally distributed across LA to cover the environmental space and we identified that efforts should be focused across Grasslands, Savanna, Open Scrublands, and Evergreen Broadleaf Forests (Table 2). The addition of new study sites progressively increased the predictive power for each model, as the area under the curve (AUC) increased from 0.003 to 0.018 and 0.003 to 0.017 for GPP and ET, respectively (Figure 6). Since the addition of 200 new sites requires an enormous logistic and economic investment, an alternative effort could be to set a goal of 60 optimally distributed sites. These sites could potentially achieve a representativeness of just below 80% of GPP and ET across LA, but with substantial uncertainty as model predictive power decreased by nearly 80% (Figure 6).

4. Discussion

4.1. Representativeness Across Land Cover, Climate, Edaphic and Topographic Characteristics

Overall, the potential spatial representativeness across LA included 34% of the surface area for climate properties, 36% for terrain parameters, 34% for soil resources, and 45% when all environmental properties were combined in multivariate space. Here we highlight the representativeness of shrublands, woody savannas, croplands and forested ecosystems (i.e., evergreen broadleaf forest and deciduous forest) as important IGBP classes across LA. Shrublands and woody savannas had the highest representativeness for environmental properties (i.e., climate, edaphic, topographic, and combined environmental properties) across all IGBP classes (Supplementary Material Table S4). This is explained because those ecosystems have a relatively high ratio of eddy-covariance sites per surface (Table 1), but their high representativeness also suggests that those ecosystems are...
Figure 3. Summaries for gross primary productivity (GPP) as GPP_mean and GPP_CV (coefficient of variation) for represented and nonrepresented FLUXNET sites (a–d). Summaries for evapotranspiration (ET) as ET_mean and ET_CV (coefficient of variation) for represented and nonrepresented FLUXNET sites (e–h).
likely less environmentally heterogeneous, arguably due to: (a) limited water conditions/stress that trigger common strategies such as water-use efficiency (Biederman et al., 2016; Huxman et al., 2004; Ponce Campos et al., 2013); and (b) a high convergence on functional and structural properties such as shrublands and savannas between North and South America (Paruelo et al., 1998). The functional and structural similarities between the shrublands and savannas within North and South America could provide insights about why scattered sites could contribute to relative high representativeness, even though most of these sites (i.e., representing shrublands and savannas) are located at the northern domain of LA (Figure 1). These results support the assumption that upsampling eddy-covariance data could be performed by using information from few sites representing some IGBP classes around the world (e.g., by FluxCom; Jung et al., 2019), but we highlight that this assumption may not be widely applicable for all IGBP classes under different environmental conditions (Vargas, Sonnentag, et al., 2013). Despite the relative high representativeness of shrublands and woody savannas, changing climate conditions such as droughts (Biederman et al., 2016; Villarreal et al., 2016) along with land cover changes (due to land-use change and disturbances) could decrease expected regional functional and structural similarities and consequently increase environmental heterogeneity across these IGBP classes.

The representativeness of forested ecosystems such as evergreen broadleaf forest and deciduous forest were consistently lower than shrublands and woody savannas (Table S4). These results are relevant because more than 50% of the sites are located across the evergreen broadleaf forest and deciduous forest (Table 1). These results suggest a larger heterogeneity in climatic, topographic, and soil resources that ultimately influence the lower representativeness within forested ecosystems. This could be a result of a larger precipitation gradient across forested ecosystems than for shrublands and savannas (F. S. Chapin et al., 2002), arguably more dynamic cycling of soil nutrients (Chapin et al., 1986; Vitousek, 1984), and larger variability in canopy phenology (Richardson et al., 2013). Improvements are needed for monitoring the large range of environmental properties of forested ecosystems across LA to properly assess and forecast the influence of global environmental change within FLUXNET sites (D. D. Baldocchi, 2020; Keenan et al., 2014). We recognize that multiple efforts on the tropical wet forests have provided important information and have fostered our knowledge on carbon and water fluxes at natural, converted, and afforested sites across LA (Andreae et al., 2002; Avissar et al., 2002; Keller et al., 2004). However, we highlight that the scientific community should be cautious when extrapolating information from a few forested sites within LA or assuming that other forested sites across the world accurately represent characteristics within LA.

Other ecosystems such as croplands and cropland/natural-vegetation mosaics have a relatively similar representativeness (as forested ecosystems) among environmental properties despite a substantially smaller number of available sites (Table 1). This lower representativeness influences the development and benchmarking of terrestrial ecosystem models as they usually have lower performance across croplands (Schaepfer et al., 2012). This limitation may hinder our capacity to forecast the influence of global environmental change on food production and food security within LA (Graesser et al., 2015; Ramankutty et al., 2002). Since disturbances such as droughts and floods are expected to cause large damage to agriculture in developing countries, a well-monitored program that allows collect-

Figure 4. Spatial representativeness across Latin America for different environmental parameters based on random forest models (a, c, e, and g), and the environmental space represented by two most influential variables for each set of environmental parameters (b, d, f, and h). Detailed statistics describing differences between represented and nonrepresented areas are in Table S4.
ing information on the impact of global environmental change on food production is imperative in order to design adaptation strategies (Graesser et al., 2015; Ramankutty et al., 2002; Reyer et al., 2015).

Other IGBP classes such as mixed forest, grasslands, and permanent wetlands have also been recognized for their important role in climate regulation and soil nutrient cycling (D. D. Baldocchi et al., 2000; Conant et al., 2001; Whiting & Chanton, 2001). Unfortunately, none of these IGBP classes were represented in our analysis (Table 1) and consequently highlight the potential bias of data-driven products when using infor-

Figure 5. Spatial representativeness across Latin America for gross primary productivity (GPP) as GPP_mean and GPP_CV (coefficient of variation; a and b). Spatial representativeness across Latin America for evapotranspiration (ET) as ET_mean and ET_CV (coefficient of variation; c and d).
This challenge is also observed for the recently compiled FLUXNET-CH4 data set where there is minimal information from tropical wetlands (despite their importance for the global methane cycle; Poulter et al., 2017) and very few sites from LA (Knox et al., 2019).

4.2. Representativeness for GPP and ET

There was a 48% potential representativeness for GPP and 34% for ET considering 41 eddy-covariance sites distributed across LA. These results compare with the representativeness by the MexFlux network of 3% for GPP and 5% for ET across Mexico (Villarreal et al., 2019), or the representativeness by AmeriFlux sites of 46% of the spatial functional heterogeneity (i.e., enhanced vegetation index dynamics) across the conterminous United States (Villarreal et al., 2018). Efforts to assess the representativeness of FLUXNET have focused on testing if measurements taken at specific locations can be extrapolated at explicit spatial and temporal extent (Chu et al., 2017; He et al., 2015; Yang et al., 2008), broadly concluding that representativeness is largely dependent on the heterogeneity of fine-scale ecosystem processes (Chen et al., 2011, 2012; He et al., 2015). Those studies highlight the challenges involved in representing regional-to-global information across heterogeneous regions (e.g., Mexico with large environmental gradients) and across larger regions even with a high density of eddy covariance sites (e.g., the United States and Europe; Sulkava et al., 2011; Villarreal et al., 2018).

Our result show that large areas of Brazil and the southern part of LA are not represented for GPP and ET. We highlight that biomes such as

| IGBP class                  | Registered sites | Including new sites |
|-----------------------------|------------------|---------------------|
| Water                       | 0                | 0                   |
| Evergreen needle-leaf forest | 0                | 0                   |
| Evergreen broad-leaf forest  | 19               | 80                  |
| Deciduous needle-leaf forest | 0                | 0                   |
| Deciduous broad-leaf forest  | 2                | 6                   |
| Mixed Forest                | 0                | 4                   |
| Closed Shrubland            | 1                | 1                   |
| Open Shrubland              | 5                | 30                  |
| Woody Savanna               | 8                | 15                  |
| Savanna                     | 1                | 54                  |
| Grassland                   | 0                | 19                  |
| Permanent Wetland           | 0                | 1                   |
| Cropland                    | 1                | 11                  |
| Urban and Build-Up          | 1                | 2                   |
| Cropland/natural vegetation mosaic | 3            | 9                   |
| Snow and Ice                | 0                | 1                   |
| Barren Land                 | 0                | 6                   |
the northern Amazon Forest, Cerrado, Caatinga, Pampa, and Pantanal in South America are not properly represented. We are aware of several eddy-covariance sites in these biomes, but they were not registered within FLUXNET nor publicly share information (at the time of our assessment) for regional-to-global data-driven products.

We highlight that our assessment of potential representativeness for GPP and ET across LA is conservative and should be seen as a current “best case” scenario if data from all sites are available for the scientific community. Consequently, most studies that have used available FLUXNET data (e.g., FLUXNET2015) for upscaling purposes, parameterize models, or performed data-syntheses that include LA are likely biased. The FLUXNET network has provided important information about how ecosystem metabolism responds to biophysical factors and key insights for improving terrestrial ecosystem process models (Baldocchi, 2020), however, its accuracy to represent these ecosystem processes at regional and global scales largely depends on the density, distribution, and diversity of sites. The addition of new sites across unrepresented regions of the world will be able to test the current assumptions of biophysical drivers for ecosystem processes and the importance of these regions for regional-to-global water and carbon cycles.

Previous studies have assessed how the distribution and density of monitoring sites affect the prediction capacity of upscaling models to properly represent GPP and ET at larger spatial scales and to capture their interannual variability (Papale et al., 2015; Sulkava et al., 2011; Zhang et al., 2020). Those results suggest that regions and/or continents with relatively few monitoring sites (e.g., Latin-America, Africa) could lead to predictions with large errors. This is also the case for modeling other ecosystem processes, such as soil respiration, where large uncertainties exist across LA and Africa due to lack of measurements to parameterize models (Warner et al., 2019). Consequently, is it inevitable to ask how accurate are data-driven products that predict ecosystem processes across regions with little or nonexistent information. This is of utmost importance for regions such as LA, since it is home to some of the largest rainforests across the globe (e.g., Amazon basin, Mayan rainforest) which are key to understand regional-to-global water and carbon cycles (Ahlström et al., 2016; Saatchi et al., 2011). Furthermore, water-limited ecosystems, which are important for the interannual variability of the global carbon cycle (Ahlström et al., 2016; Poulter et al., 2014), may have different GPP and ET dynamics as expected for other water-limited ecosystems around the world (Zhang et al., 2020).

Finally, there is an important challenge for monitoring and representing GPP and ET dynamics from the regional-to-global scales. The global representation of FLUXNET sites is biased toward undisturbed sites (Baldocchi, 2020) and LA is not the exception. Disturbances such as hurricanes (Vargas, 2012), grazing of pastures (Wolf et al., 2011) and land use (Cabral et al., 2020) are important to determine the attribution of changes in regional water and carbon budgets. Furthermore, it is critical to assess the effects of ecological degradation and subsequent ecosystem recovery to properly assess carbon and water dynamics in ecosystems across LA (Bustamante et al., 2016; Vargas et al., 2008). Examples of challenges involved with the complexity of LA landscapes include: (a) the debate on the role of the Amazon basin as a carbon sink our source (Andreae et al., 2002; Avisar et al., 2002; Davidson & Artaxo, 2004); (b) testing the consistency of the expected carbon use efficiency (the ratio between ecosystem respiration and GPP) for disturbed ecosystems (D. Baldocchi & Penuelas, 2019); or (c) how old growth tropical forests respond to weather variability (Rojas-Robles et al., 2020). Consequently, we encourage the scientific community to consider monitoring efforts across the IGBP classes highlighted in this study, but also the land use history and state of the ecosystems for better representations of the water and carbon cycles across LA and the world.

### 4.3. Influence of Spatial Scale on the Representativeness of GPP and ET

The FLUXNET database has been extensively used for the development of terrestrial ecosystem process models and data-driven products (D. D. Baldocchi, 2020; Jung et al., 2019), but there are challenges associated with: (a) the distribution of sites that provide model parameterization and ground truth data; and (b) the mismatch between the footprint represented by the ground truth data and the footprint (i.e., pixel size) from the models’ output. Here, we discuss how the representativeness of GPP and ET is influenced by different spatial scales typically used for global estimations (i.e., spatial grids with resolutions at 0.25°, 0.50°, 1.0°).
We found that as the pixel size increases (i.e., at coarser spatial resolutions) representativeness decrease to about 25% for GPP and ET across LA. Again, these results are conservative and assuming that all data from the 41 sites is available for the scientific community. One explanation about why representativeness decrease as pixel size increase is that there are study sites within close proximity that may fall within one pixel at coarser spatial resolutions. Consequently, many environmental properties are not represented by the fewer sites included within the coarser pixels; in other words, fewer pixels are available to provide information across LA. Future studies could focus on how the density of sites across LA influence modeling errors for GPP and ET (Papale et al., 2015) or upscaling data-driven products relevant for the regional (Guevara et al., 2018) and global carbon cycle (Jung et al., 2019; Warner et al., 2019). Consequently, an important issue is where to establish new study sites to improve the representativeness of FLUXNET across LA.

4.4. Improving Representativeness

LA incorporates one of the most ecologically diverse regions in the world, having a large influence on the global carbon and water cycles, the Earth’s climate system and global biogeochemical cycles (Balvanera et al., 2012). Our results show spatial representativeness gaps among assessed environmental properties (i.e., bioclimatic, terrain properties, and soil resources) including GPP and ET (Figures 1 and 2) to identify underrepresented regions from different IGBP classes (Table S4). We highlight that larger monitoring efforts should focus on the evergreen broadleaf forest, croplands, savanna, and open shrublands, but including permanent wetlands and snow and ice classes are needed to cover the full spectrum of biomes across LA (Table 2). Furthermore, we recognize that it is important to consider past land use and disturbances, but this is usually not assessed in FLUXNET synthesis studies.

What would happen if 200 strategically located sites are installed across LA? Our results show that: (a) representativeness would increase to 86% and 80% of GPP and ET, respectively (Figure 5); and (b) the correlation between represented regions and environmental properties monitored by these sites would increase. These results are based on a theoretically distributed network of sites to optimally represent the environmental space across LA. We highlight that 60 optimally distributed sites could achieve similar representativeness for GPP and ET but with higher uncertainty (Figure 5).

Previous studies have highlighted the need to increase monitoring across underrepresented ecosystems within FLUXNET, such as water-limited ecosystems (Hargrove et al., 2003; Papale et al., 2015; Villarreal et al., 2018), and it is known that increasing the number of sites within a region of interest would result in lower uncertainty estimates (Papale et al., 2015; Sulkava et al., 2011; Villarreal et al., 2019). However, there is always a tradeoff between monitoring space and time to properly represent spatial and temporal variability. Our study focuses on maximizing spatial representativeness of long-term means of environmental properties, GPP and ET, but we highlight that temporal representativeness should also be considered (Chu et al., 2017). These assessments could be tested across LA as more sites with long-term eddy-covariance records are included and compared across our understanding of the temporal and spatial representativeness of the FLUXNET network.

4.5. Challenges and Opportunities

The overall scope of this study was to provide an assessment of the representativeness of registered FLUXNET sites within LA to promote monitoring efforts and sharing of information among local, regional and global networks. We highlight that many of the registered FLUXNET sites have not shared information with FLUXNET or AmeriFlux and consequently their data is not widely available for the scientific community to perform synthesis studies and data-driven products. By hence, our representativeness analysis must be taken as a conservative approach of a “best case” scenario, especially for those IGBP classes with no registered eddy-covariance sites (Table 1). For this study, we assumed that if a site was registered with FLUXNET by 2018, then the eddy-covariance information either is available, or the principal investigator is willing to contribute with FLUXNET in the near future. A clear example are the sites located across Mexico (i.e., MexFlux; Vargas et al., 2013) that are affiliated with FLUXNET but most data are not currently available for the wider scientific community (Villarreal et al., 2019). We are also aware that there are multiple unaffiliated sites across LA such as in semi-arid grasslands (Hinojo-Hinojo et al., 2016), croplands (Lewczuk et al., 2017),
The Chacon region (Toledo & Figuerola, 2019), The Andes (Callañaupa-Gutierrez et al., 2020), Patagonia (Valdés-Barrera et al., 2019), among other sites across LA. Furthermore, this study did not account for representation of disturbances and land use change across LA. Including sites across the Caribbean islands could also improve the representativeness across tropical regions that are highly affected by tropical storms and hurricanes. We hope that this study motivates principal investigators and regional networks (i.e., Mex-Flux, Brasflux, or SULFLUX) to join and contribute to FLUXNET to build a stronger global network and increase our understanding of LA in the global water and carbon cycles.

Traditionally, the establishment of eddy-covariance sites across LA has not been done under a national coordinated effort, and consequently, monitoring efforts have been performed by individual research groups or by local networks with clear questions focused on specific biomes (Robert et al., 2012; Vargas et al., 2013). Furthermore, increasing the representativeness of FLUXNET across LA, and other areas of the world such as Africa, is difficult for several reasons. There are challenges related to limited economic resources (e.g., limited funding opportunities, increased costs due to importation taxes), human resources (e.g., fewer trained scientists to operate eddy-covariance sites and analyze data), and logistic issues (e.g., security and accessibility).

Our results bring attention to the possibilities that exist by coordinating optimized monitoring efforts to improve FLUXNET spatial representativeness. Although adding 60–200 new study sites is currently not possible, there are a few initial steps that could be done: (a) a coordinated effort to archive, synthesized and analyze existing information across LA; (b) collaboration on the collection of new information across underrepresented sites with national, regional or partnerships outside LA; and (c) promote a stronger culture of data sharing and proper recognition of the efforts of researchers across LA. We recognize that these efforts will require an increase of interoperability across existing networks and researchers within LA, and we must work as a community to reduce conceptual, technological, organizational, and cultural barriers (Vargas et al., 2017). Finally, we warn about “helicopter research,” which can be understood as research mostly performed by institutions/researchers from developed countries within developing countries with lesser involvement from local researchers (Minasny et al., 2020), and call for quality and equality of partnerships where there should be a mutual benefit between local scientists (with local infrastructure and local knowledge) and foreign scientists performing research across LA. We advocate for data sharing following FAIR (i.e., findable, accessible, interoperable, and reproducible) data principles (A. A. Fisher & Koven, 2020), but at the same time strongly support that contributions from LA scientists should be properly recognized and supported by the FLUXNET community. A positive relationship between the wider scientific community and scientist within LA will encourage partnerships among institutions and will promote local contributions to increase the representativeness of information for regional-to-global water and carbon budgets.

5. Conclusions

The present study used machine learning methods to assess the representativeness of FLUXNET sites across LA as it has been applied to previous studies assessing the representativeness of EONs (Villarreal et al. 2018, 2019). This study provides an overall scope of FLUXNET representativeness gaps of climate, topographic, soil resources along with GPP and ET across LA, a region of major importance for global biogeochemical cycles. With 41 sites registered in FLUXNET (revised in 2018 but see methods for details) across LA, our results show that in a best-case scenario, spatial representativeness of environmental properties such as climate, topography, and soil resource is around 40%, while for GPP and ET the spatial representativeness is 48% and 34%, respectively. These results bring attention to evaluate the representativeness and accuracy of global data-driven products of ecosystem processes.

The overall representativeness of FLUXNET could substantially increase if 60–200 sites were optimally located across LA to properly represent the environmental space of the region. We highlight that efforts could be focused in evergreen broadleaf forest, savanna, and open shrublands; however, to coordinate such an effort there should be a higher degree of interoperability among scientists, research groups, and local-to-global networks. The benefits from enhancing FLUXNET representativeness across LA will help to improve our knowledge on the impact of global environmental change and support science-based environmental public policies and management decisions.
Data Availability Statement

Data sets used to perform this research are publicly available in the following data repositories: climate variables (https://worldclim.org/); soil organic carbon (http://www.fao.org/soils-portal/data-hub/soil-maps-and-databases/global-soil-organic-carbon-map-gsocmap/en/); soil phosphorus (https://data.nasa.gov/dataset/A-Global-Database-of-Soil-Phosphorus-Compiled-from/9kgt-96rs); MOD17A2 (https://lpdaac.usgs.gov/products/mod17a2v006/); MOD16A2 (https://lpdaac.usgs.gov/products/mod16a2v006/).

Acknowledgments

Samuel Villarreal acknowledge support from CONACyT doctoral fellowships. Rodrigo Vargas acknowledges support from NASA Carbon Monitoring Systems (80NSSC18K0173). The authors thank researchers and scientists from Latin America for their contributions to better understand land-atmosphere interactions and the broader FLUXNET community for inspiring this study.

References

Ahlström, A., Raupach, M. R., Schurgers, G., Smith, B., Arnette, A., Jung, M., et al. (2016). The dominant role of semi-arid ecosystems in the trend and variability of the land CO₂ sink. Science, 348(6237), 895–899. https://doi.org/10.1126/science.1251022
Alcaraz-Segura, D., Lomba, A., Sousa-Silva, R., Nieto-Lugilde, D., Alves, P., Georges, D., et al. (2017). Potential of satellite-derived ecosystem functional attributes to anticipate species range shifts. International Journal of Applied Earth Observation and Geoinformation, 57, 86–92. https://doi.org/10.1016/j.jag.2016.12.009
Amundson, R. J. H. (1991). On a state factor model of ecosystems. Biogeochemistry, 47(8), 536–543.
Andere, M. O., Artaxo, P., Brandão, C., Carvalho, F. E., Cicioli, P., Da Costa, A. L., et al. (2002). Biogeochemical cycling of carbon, water, energy, trace gases, and aerosols in Amazonia: The LBA-EUSTACH experiments. Journal of Geophysical Research D, 107(D20). https://doi.org/10.1029/2001JD000524
Avissar, R., Dias-Silva, P. L., Silva-Dias, M. A. F., & Nobre, C. (2002). The Large-Scale Biosphere-Atmosphere Experiment in Amazonia (LBA): Insights and future research needs. Journal of Geophysical Research, 107(D20), 8086. https://doi.org/10.1029/2002JD002704
Baldocchi, D. D. (2020). How eddy covariance flux measurements have contributed to our understanding of Global Change Biology. Global Change Biology, 26, 242–260. https://doi.org/10.1111/gcb.14807
Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., 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, 2415–2435.
Baldocchi, D., & Penuelas, J. (2019). The physics and ecology of mineral carbon dioxide from the atmosphere through ecosystem services. Global Change Biology, 25(4), 1191–1197. https://doi.org/10.1111/gcb.14559
Baldocchi, D. D., Finnigan, J., Wilson, K. B., U. T. P., & Falge, E. (2000). On measuring net ecosystem carbon exchange over tall vegetation on complex terrain. Boundary-Layer Meteorology, 96(1–2), 257–291. https://doi.org/10.1023/A:1002497615457
Balvanera, P., Uriarte, M., Almeida-Leñero, L., Altester, A., DeClerck, F., Gardner, T., et al. (2012). Ecosystem services research in Latin America: The state of the art. Ecosystem Services, 2, 56–70. https://doi.org/10.1016/j.ecoser.2012.09.006
Barbet-Massin, M., Jiguet, F., Helene Albert, C., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: how, where and how many?. Methods in Ecology and Evolution, 3, 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x
Biederman, J. A., Scott, R. L., Goulden, M. L., Vargas, R., Litvak, M. E., Kohl, T. E., et al. (2016). Terrestrial carbon balance in a drier world: The effects of water availability in southwestern North America. Global Change Biology, 22(5), 1867–1879. https://doi.org/10.1111/gcb.13222
Brienen, R. J., Phillips, O. L., Feldpausch, T. R., Gloor, E., Baker, T. R., Lloyd, J., et al. (2015). Long-term decline of the Amazon carbon sink. Nature, 519(7543), 344–348.
Bustamante, M. C., Roitman, I., Aide, T. M., Alencar, A., Anderson, L. O., Aragão, L., et al. (2016). Toward an integrated monitoring framework to assess the effects of tropical forest degradation and recovery on carbon stocks and biodiversity. Global Change Biology, 22(1), 92–109. https://doi.org/10.1111/gcb.13087
Cabrál, O. M. R., Freitas, H. C., Tye, M. R., & Hudnut, D. (2017). Integrated experimental research infrastructure: A paradigm shift to face an uncertain world and innovate for societal benefit. CRC Press.
Chabbi, A., Loescher, H. W., Tye, M. R., & Hudnut, D. (2017). Integrated experimental research infrastructure: A paradigm shift to face an uncertain world and innovate for societal benefit. CRC Press.
Chapin, F. S., Matson, P. A., & Mooney, H. A. (2002). Principles of Terrestrial Ecosystem Ecology. System, 21. https://doi.org/10.1111/0075-5435.00494
Chapin, F. S., Vitousek, P. M., & Van Cleve, K. (1986). The nature of nutrient limitation in plant communities. The American Naturalist, 127(1), 48–58. https://doi.org/10.1086/244466
Chen, B., Coops, N. C., Fu, D., Margolis, H. A., Amiro, B. D., Barr, A. G., et al. (2011). Assessing eddy-covariance flux tower location bias across the Fluxnet-Canada Research Network based on remote sensing and footprint modelling. Agricultural and Forest Meteorology, 151(1), 87–100. https://doi.org/10.1016/j.agrformet.2010.09.005
Chen, B., Coops, N. C., Fu, D., Margolis, H. A., Amiro, B. D., Black, T. A., et al. (2012). Characterizing spatial representativeness of flux tower eddy-covariance measurements across the Canadian Carbon Program Network using remote sensing and footprint analysis. Remote Sensing of Environment, 124, 742–755. https://doi.org/10.1016/j.rse.2012.06.007
Chrysoulakis, N., Spiliotopoulos, M., Domenikiotis, C., & Dalezios, N. (2003). Towards monitoring of regional atmospheric instability through modis/aqua images: In: Proceedings of the International Symposium held at Volos, November, Greece.
Chu, H., Baldocchi, D. D., John, R., Wolf, S., & Reichstein, M. (2017). Fluxes all of the time? A primer on the temporal representativeness of FLUXNET Housen. Journal of Geophysical Research: Biogeosciences, 122, 289–307. https://doi.org/10.1002/2016JG003576
Conant, R. T., Faustian, K., & Elliott, E. T. (2001). Grassland management and conversion into gruassland: Effect on soil carbon. Ecological Applications, 11(2), 343–355. https://doi.org/10.1890/1051-0761(2001)011[0343:GMACBG]2.0.CO;2
Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. Ecology, 88(11), 2783–2792. https://doi.org/10.1890/07-0539.1
