Are Remote Sensing Evapotranspiration Models Reliable Across South American Ecoregions?


1Federal University of Paraíba, Areia, PB, Brazil, 2Federal University of Mato Grosso do Sul, Campo Grande, MS, Brazil, 3Hydro-Climate Extremes Lab (H-CEL), Ghent University, Ghent, Belgium, 4Schmid College of Science and Technology, Chapman University, Orange, CA, USA, 5Department of Life Sciences, Imperial College London, Silwood Park Campus, Ascot, UK, 6Numerical Terradynamic Simulation Group, University of Montana, Missoula, MT, USA, 7Brazilian Agricultural Research Corporation, Embrapa Meio Ambiente, Jaguariúna, SP, Brazil, 8Department of Atmospheric and Climate Sciences, Federal University of Rio Grande do Norte, Natal, RN, Brazil, 9Climate Sciences Graduate Program, Federal University of Rio Grande do Norte, Natal, RN, Brazil, 10Department of Hydrology and Atmospheric Sciences, The University of Arizona, Tucson, AZ, USA, 11Brazilian Agricultural Research Corporation — Embrapa Tropical Semi-árid, Petrolina, PE, Brazil, 12Federal University of Mato Grosso, Cuiabá, MT, Brazil, 13University of Bristol, Bristol, UK, 14Department of Biological and Agricultural Engineering, Texas A&M University, College Station, TX, USA, 15Department of Nuclear Energy, Federal University of Pernambuco, Recife, PE, Brazil, 16Center for Earth System Research and Sustainability (CENS), University of Hamburg, Hamburg, Germany, 17Department of Environmental Science and Renewable Natural Resources, University of Chile, Santiago, Chile, 18Institute of Ecology and Biodiversity, Santiago, Chile, 19Department of Geography and Environmental Science, The University of Reading, Reading, UK, 20Federal University of the Agreste of Pernambuco, Garanhuns, PE, Brazil, 21Federal Rural University of Pernambuco, Serra Talhada, PE, Brazil, 22Department of Atmospheric and Ocean Sciences, FCEN — UBA, Buenos Aires, Argentina, 23National Council for Scientific and Technical Research, CONICET, Buenos Aires, Argentina, 24Instituto de Clima y Agua. Instituto Nacional de Tecnología Agropecuaria (INTA), Hurlingham, Argentina, 25Department of Hydraulics and Sanitary Engineering, University of São Paulo, São Carlos, SP, Brazil

Abstract Many remote sensing-based evapotranspiration (RSBET) algorithms have been proposed in the past decades and evaluated using flux tower data, mainly over North America and Europe. Model evaluation across South America has been done locally or using only a single algorithm at a time. Here, we provide the first evaluation of multiple RSBET models, at a daily scale, across a wide variety of biomes, climate zones, and land uses in South America. We used meteorological data from 25 flux towers to force four RSBET models: Priestley–Taylor Jet Propulsion Laboratory (PT-JPL), Global Land Evaporation Amsterdam Model (GLEAM), Penman–Monteith Mu model (PM-MOD), and Penman–Monteith Nagler model (PM-VI). ET was predicted satisfactorily by all four models, with correlations consistently higher ($R^2 > 0.6$) for GLEAM and PT-JPL, and PM-MOD and PM-VI presenting overall better responses in terms of percent bias ($-10 < PBIAS < 10\%$). As for PM-VI, this outcome is expected, given that the model requires calibration with local data. Model skill seems to be unrelated to land-use but instead presented some dependency on biome and climate, with the models producing the best results for wet to moderately wet environments. Our findings show the suitability of individual models for a number of combinations of land cover types, biomes, and climates. At the same time, no model outperformed the others for all conditions, which emphasizes the need for adapting individual algorithms to take into account intrinsic characteristics of climates and ecosystems in South America.

1. Introduction

Land evaporation, or evapotranspiration (ET), is the phenomenon by which water is converted from a liquid into its vapor phase over land. It plays a significant role in the modulation of global climate feedbacks...
being a key driver of the Earth’s carbon, energy, and water cycles at local, regional, and global scales (Cao et al., 2010; de Oliveira et al., 2021; Khosa et al., 2019; Tong et al., 2017; Valle Júnior et al., 2020). In situ ET measurements can be obtained from micro-meteorological methods (e.g., eddy covariance, scintillometry, or Bowen ratio method) and those derived from the soil water balance (e.g., directly using lysimeters, or from changes in profile soil moisture content obtained gravimetrically, from neutron probes, or capacitance-based soil water monitoring equipment). Besides, plant physiological techniques such as sap flow methods, provide direct estimates of transpiration (Allen et al., 2011; Fisher et al., 2011; Verhoef & Campbell, 2006), but only the micro-meteorological methods provide ET data at the field to landscape (e.g., scintillometry) scale. Over the past three decades, eddy covariance (EC) systems have become the state-of-the-art and standard in situ method to quantify land surface energy and mass fluxes for different types of ecosystems (Campos et al., 2019; Restrepo-Coupe et al., 2013; Rodrigues et al., 2016; Wang et al., 2020). However, these techniques estimate fluxes for areas of relatively limited spatial dimensions (~1 km²) depending on the heterogeneity of the landscape, and they are affected by specific local conditions, such as the occurrence of advection across sharp contrasts in vegetation and/or irrigation conditions, and those caused by topographic features, such as cold air drainage for sloping terrain (Allen et al., 2011; Mauder et al., 2020; Mutti et al., 2019; Rahimzadegan & Janani, 2019; Rwasoka et al., 2011).

During the 1990s and 2000s, remote sensing-based ET (RSBET) algorithms, using information from visible, near-infrared, and thermal infrared bands, were developed, such as the Surface Energy Balance Algorithms for Land (SEBAL, Bastiaanssen et al., 1998), Simplified Surface Energy Balance Index (S-SEBI, Roerink et al., 2000), Surface Balance Energy System (SEBS, Su, 2002), Simplified Surface Energy Balance (SSEB, Senay et al., 2007), and Two-Source Energy Balance Model (TSEB, Kustas & Norman, 1999; Norman et al., 1995). These algorithms were developed for sub-regional applications, with a focus on irrigation or water resources management. Over South America, their predictive skills have been assessed extensively, mostly for irrigated cropland (Bezerra et al., 2013, 2015; Lopes et al., 2019; Mutti et al., 2019; Oliveira-Guerra et al., 2017; Paiva et al., 2011; Poblete-Echeverría & Ortega-Farias, 2012; Teixeira et al., 2009). Studies show that these models perform well when compared to field observations of ET (Poblete-Echeverría & Ortega-Farias, 2012; Teixeira et al., 2009).

Since the late 2000s, algorithms such as PT-JPL (Fisher et al., 2008), PM-MOD (Mu et al., 2007, 2011), and GLEAM (Martens et al., 2017; Miralles et al., 2011) focused on the use of satellite-derived observations to create spatially coherent global ET estimates (Fisher et al., 2017). PT-JPL is at the core of the ECOSTRESS mission (Fisher et al., 2020), while PM-MOD is central to the global terrestrial MODIS ET product (MOD16). GLEAM is used for the annual State of the Climate report since 2015 (Blunden & Arndt, 2020).

Using flux tower data, previous studies conducted in South America evaluated GLEAM and MOD16 (Moreira et al., 2019; Paca et al., 2019; Ruhoff et al., 2013). However, these studies validated off-the-shelf ET datasets generated by these models, not the models themselves. Since such ET products are not produced using a common dataset of meteorological variables, a comparative evaluation cannot be made in terms of model structure. Rather, different model skills would be partially linked with the quality of the inputs. A multisite tropical study, over several continents, validating the PT-JPL model at a regional scale on a monthly basis was presented by Fisher et al. (2009). However, to the best of our knowledge, studies assessing the daily predictive skills have only been conducted at the local scale (Miranda et al., 2017; Oliveira et al., 2018; Souza et al., 2019; Teixeira et al., 2009, 2013).

A major challenge to verify the results of these methods is the scarcity of ground-based observations, due to the uneven spatio-temporal distribution of the ET monitoring efforts. As a result, remote sensing ET methods are typically evaluated or parameterized using sites located only in North America, Europe (Ershadi et al., 2014; McCabe et al., 2016; Michel et al., 2016; Xu et al., 2019), Australia (Martens et al., 2016), and East Asia (Chang et al., 2018; Jang et al., 2013; Khan et al., 2018; Li et al., 2019). For example, Mu et al. (2011) proposed improvements to the PM-MOD ET global algorithm (Mu et al., 2007), based on comparisons with ET measurements from 46 AmeriFlux sites, 45 of them located in USA and Canada. Martens et al. (2017) evaluated the GLEAM algorithm with 91 worldwide FLUXNET sites; however, ~65 were located in the USA and in Europe. Therefore, these models might not satisfactorily represent ET in sparsely sampled regions with very different climate conditions such as South America, despite this continent representing ca. 12% of the total Earth’s terrestrial area.
South America spans two hemispheres, and four major climate zones, from near the equator to sub-Antarctic regions, which makes it a geographically unique continent (Goymer, 2017; Trajano, 2019). Biomes in this continent range from tropical to deciduous forests, and contain ecoregions with high sensitivity to variability in water (e.g., the Caatinga and Humid Pampas) and energy availability (e.g., the Amazon, Valdivian temperate, and Magellanic subpolar forests) (Seddon et al., 2016). Also, five out of six of the terrestrial biomes not included in satellite-based ET algorithm evaluations at a global scale are found in South America (see Section 2.1). Thus, the evaluation of RSBET methods for South America offers an opportunity to reduce the current research gap, in particular at large spatial scales.

FLUXNET provides a common framework for the verification of ET algorithms. Nevertheless, the available sites in the FLUXNET2015 database are not evenly distributed around the world (Pastorello et al., 2020). Validating global models in South America is challenging, mainly because the data from ~90% of its FLUXNET registered sites are not readily available to the scientific community: less than 50% of South American AmeriFlux sites are available for direct access. Additionally, flux towers in woody savannas and evergreen broad-leaf forests account for nearly 65% of all Latin American FLUXNET sites, while some of the biomes are not properly represented (Villarreal & Vargas, 2021).

The identification of scientific gaps and the proposed improvements are considered a priority for the future development of ET assessment methods from remote sensing (Fisher et al., 2017). Some of them include merging different ET-estimation methods and the identification of their sources of uncertainty (Fisher et al., 2017; Paca et al., 2019; Zhang et al., 2017). Indeed, despite the recent developments of remote sensing ET methods, there are still challenges concerning the refinement of those algorithms to remedy the lack of information on specific surface characteristics and fluxes of undersampled climate zones and vegetation types. In this context, one of the main sources of uncertainty in global satellite-based ET estimates are the fractional vegetation cover and net radiation (Badgley et al., 2015; Ferguson et al., 2010; Vinukollu et al., 2011).

We evaluated the predictive skills of four satellite-based ET models designed for regional and continental scale applications, over South America. The main question we seek to answer is whether such models can be applied consistently to reliably capture ET in South America. Specific research questions include: (a) are the models capable of correctly estimating ET and its components? (b) are the models predictive skills affected by climate, land cover type, or biome?

2. Study Area, Data, and Methods

2.1. South American Biomes, Flux Tower-Based ET, and Meteorological Data

The study area encompasses five biomes (Table S1 in the Supporting Information S1): Tropical and Subtropical Moist Broadleaf Forests (TSMBF); Flooded Grasslands and Savannas (FGS); Tropical and Subtropical Grasslands, Savannas and Shrublands (TSGSS); Tropical and Subtropical Dry Broadleaf Forests (TSDBF), and Temperate Broadleaf & Mixed Forests (TBMF) (Olson et al., 2001).

We used daily meteorological data from 25 flux tower sites located across various South American biomes and land cover types to verify the predictive skill of the selected RSBET models (Figure 1a, Table S2 in Supporting Information S1). The time period considered for analysis was determined by the available time series for each site (Figure S1 in Supporting Information S1). Further information about each biome is provided in SM. 10 sites are from FLUXNET (Pastorello et al., 2020), AmeriFlux networks (Novick et al., 2018), and Large-Scale Biosphere-Atmosphere Experiment in the Amazon (LBA) project (Saleska et al., 2013), while the remaining data were obtained from the respective principal investigators. Concerning towers sites not available in global networks, data handling included standard procedures to ensure quality data, including: detection of spikes caused by changes in the footprint or imprecise measurements; delay correction of H₂O/CO₂ in relation to the vertical wind component; correction of coordinates (2D rotation); correction of spectral loss; conversion of the buoyancy flux to sensible heat flux, known as SND-corrections (Schotanus et al., 1983); sonic virtual temperature correction; corrections for flux density fluctuations, known as WPL corrections (Webb et al., 1980); and incorporated frequency response correction. Additionally, we performed due corrections with respect to reduction of wind velocity or turbulence increase caused by the shadow of the tower and the sensor. Details about procedures carried out for data processing and filtering.
to implement these corrections can be found in Cabral et al. (2020); Campos et al. (2019); Holl et al. (2019); Tonti et al. (2018). We also emphasize that those data have been widely used in previously scientific publications (Arruda et al., 2016; Cabral et al., 2010, 2011; Marques et al., 2020; Restrepo-Coupe et al., 2013; Rocha et al., 2009; Rodrigues et al., 2016; Silva et al., 2017). The spatial patterns of mean annual precipitation ($P$), air temperature ($T$), and potential evapotranspiration ($PET$) show that selected sites encompass a wide variety of climates (Figure 1b).

The closure of the energy budget (EB) is rarely observed with flux tower measurements (Foken, 2008; Wilson et al., 2002). Usually, the available energy flux ($Rn - G$) is greater than ($LE + H$), where $Rn$ is the net radiation, $G$ is the soil heat flux, $LE$ is the latent heat flux, and $H$ is the sensible heat flux. The imbalances...
in the surface energy balance, reported here as an energy balance ratio, EBR (i.e., \((LE + H)/(Rn - G))\), range from 0.73 to 1.16 (mean \(~0.90\) (Table S2 in Supporting Information S1). It is paramount that only high-quality data were used to run and assess the models. We computed daily EBR for each site and excluded days with EBR \(\leq 0.75\) or \(\geq 1.25\). Daily averages of meteorological variables were calculated from 30-min or hourly data only when at least 80% of the records per day were available. That same criterion was applied to sub-daily data when obtaining daytime and nighttime inputs for the MOD16 model (PM-MOD in this paper), we considered only days with a minimum of 20–30-min daytime records and 20 during the night. As in Mu et al. (2011), the shortwave incoming radiation \((Rgs \downarrow)\) was used to distinguish between daytime \((Rgs \downarrow >10 \text{ W m}^{-2})\) and nighttime \((Rgs \downarrow < 10 \text{ W m}^{-2})\). Regarding the fluxes, we used quality-checked data that had not been gap-filled. Previous studies have shown that ET derived from the other energy balance fluxes, that is, \(LE = Rn - G - H\), can agree well with eddy covariance ET and lysimeter data (Amiro, 2009; Sánchez et al., 2019). Therefore, instead of using the EC-measured \(LE\), to represent ET, we derived LE from the equation above. Such a validation approach (i.e., comparing model \(ET\) with EB-derived \(LE\), \(ET_{EB}\)) has been adopted in previous studies (Fisher et al., 2020; Stoy et al., 2013; Twine et al., 2000; Wilson et al., 2002). The results of using the eddy covariance \(ET\) \((ET_{EC})\) instead can be found in the Supporting Information S1 (see Figures S11–S13 in Supporting Information S1).

The quality control procedure described above was not adopted for the TF1 and TF2 towers (see Figure 1a). At those sites, horizontal advection plays an important role due to extreme weather variations throughout the year (Levy et al., 2020), such that the energy balance closure cannot be diagnosed by EBR, as described above. For instance, the SDF zone is known as an anticyclone pathway between the Pacific and Atlantic oceans, and TF1 and TF2 are located in the extreme southern parts of Patagonia, a region characterized by strong winds. Thus, for TF1 and TF2 sites, we used \(ET\) derived from measured \(LE\).

### 2.2. Remote Sensing-Based Vegetation Indices

The required vegetation index (VI) to run PT-JPL, PM-MOD, and PM-VI is the Enhanced Vegetation Index (EVI). Vegetation Optical Depth (VOD) is used in GLEAM. EVI was derived from the 16-day Level 3 Global product of the MODerate Resolution Imaging Spectroradiometer (MODIS), aboard the Terra and Aqua satellites (Huete et al., 2002). We used both MODIS VI products, that is, MOD13Q1 (Terra) and MYD13Q1 (Aqua), at 250 m resolution, to derive daily composites of EVI. VOD was extracted from the product described in Moesinger et al. (2020). Fisher et al. (2008) used the Soil Adjusted Vegetation Index (SAVI) instead of EVI because the former does not require the blue reflectance (0.45–0.51 \(\mu\)m), however, the authors recognize that both indices are very similar. As we are interested in assessing the \(ET\) models rather than the products resulting from different forcing data, we used EVI in Fisher’s model (PT-JPL). Leaf area index (LAI) and other vegetation-related variables (e.g., fraction of Absorbed Photosynthetically Active Radiation, \(f_{PAR}\)) are handled differently in each model. For example, in PT-JPL, LAI is obtained from total fractional vegetation cover, whereas in PM-MOD the 1-km MODIS LAI (MOD15) product is adopted. The original procedures to obtain those variables were not changed here. The following treatment was applied to the MODIS-derived data. “Good quality” pixels were selected, based on the quality assurance (QA) flags. Next, an autoregressive model was applied to fill in the gaps (Akaike, 1969). The gap-filling procedure was applied to gaps smaller than 16 days, while gaps of longer periods were excluded from the analysis. Finally, we implemented a temporal filter to improve the \(f_{PAR}\) and LAI time series to reproduce precisely all preprocessing steps of the standard PM-MOD algorithm (Mu et al., 2011). Filtering of \(f_{PAR}\) and LAI allowed for the correction of underestimated values (abrupt and unrealistic decreases in the time series) that mostly originate from cloud contamination effects, which were not correctly identified in the quality control fields.

### 2.3. Summary of Remote Sensing-Based ET Models

#### 2.3.1. GLEAM

GLEAM is a semi-empirical/process-based model that estimates the total evaporative flux and its components. In this study, version 3 of the algorithm is used (Martens et al., 2017). The main aspects of the model are described briefly, while for details we refer to Miralles et al. (2011) and Martens et al. (2017). The model
calculates potential evaporation for four sub-grid land cover fractions: (a) open water, (b) low vegetation, (c) tall vegetation, and (d) bare soil using the Priestley and Taylor (1972) equation. For tall and low vegetation cover fractions, potential transpiration is constrained using an empirical evaporative stress factor, which is calculated as a function of soil moisture at root-zone depth and microwave VOD as described in Martens et al. (2017). VOD (Vegetation Optical Depth) accounts for the attenuation of microwaves through vegetation and can be used as a proxy for vegetation phenology. Thus, VOD is a microwave parameter closely linked to vegetation water content (Liu et al., 2013) and in GLEAM it is used to represent phenological changes in vegetation. The soil moisture in the root-zone is calculated with a multi-layer water-balance model forced by precipitation and satellite surface soil moisture retrievals. For bare soil, the evaporative stress factor is calculated as a function of surface soil moisture only, whereas for open water evaporation, no stress factor is applied. For the tall vegetation cover fraction, rainfall interception loss is estimated with Gash's analytical model (Gash, 1979; Miralles et al., 2010; Valente et al., 1997). The ET is then calculated as the sum of low and tall vegetation transpiration, rainfall interception loss, bare soil evaporation, and open-water evaporation with each weighted by the respective fraction.

2.3.2. PT-JPL

The global ET model proposed by Fisher et al. (2008) is based on the Priestley and Taylor equation for potential ET, which is partitioned into actual plant transpiration, soil evaporation, and interception evaporation, that is, \( E_{\text{trans}} + E_{\text{soil}} + E_{\text{int}} \). To reduce potential ET to actual ET, the PT-JPL model applies ecophysiological constraints based on land surface information such as vegetation properties and humidity/water vapor pressure deficit (VPD). Fisher et al. (2008) used NDVI and SAVI as a proxy for plant physiological status. We used EVI because it provides a better indication of green vegetation cover than NDVI, as acknowledged by Fisher et al. (2008). The model partitions available energy flux using four plant-related constraints: LAI, green canopy fraction, plant temperature, and plant moisture. Similar to PM-MOD (see next subsection), vegetation cover, canopy wetness, etc. determine how the available energy flux is partitioned among the ET terms. A unique aspect related to the plant temperature constraint is the determination of an optimal temperature, \( T_{\text{opt}} \) (Potter et al., 1993), which corresponds to an optimal stomatal conductance. The latter codetermines \( E_{\text{trans}} \).

2.3.3. PM-MOD

The MOD16 ET model (PM-MOD) is based on the Penman-Monteith equation to produce a daily global ET product summing up daytime and nighttime ET (Mu et al., 2011). In this model, total ET is partitioned into \( E_{\text{soil}}, E_{\text{int}}, \) and \( E_{\text{trans}} \). To compute \( E_{\text{soil}} \), PM-MOD uses potential soil evaporation and a soil moisture constraint function based on VPD and air relative humidity (RH) (Fisher et al., 2008). The evaporation of the water intercepted by the canopy, \( E_{\text{int}} \), is calculated using the relevant equations from a revised version of the Biome-BGC model (Thornton, 1998). The PM-MOD assumes that \( E_{\text{soil}} \) occurs when the vegetation is covered with water, i.e., when the water cover fraction \( f_{\text{soil}} > 0 \), which is constrained by RH (Mu et al., 2011). In the PM-MOD model, \( f_{\text{soil}} \) is calculated as in the PT-JPL model: \( f_{\text{soil}} \) is set to 0 if \( RH < 70\% \) and \( f_{\text{soil}} = (\text{RH}/100)^4 \) if \( 70 < RH < 100\% \) (Running et al., 2019). The PM-MOD model is designed to allow \( E_{\text{trans}} \) to occur during daytime and nighttime, by adding constraints to stomatal conductance for VPD and minimum air temperature, and ignoring constraints relating to high air temperature (Running et al., 2019). The partitioning of available energy flux into soil or interception evaporation is based on vegetation cover \( f_C \), which is assumed to be equal to \( f_{\text{PAR}} \) from the MODIS product MOD15A2 (Mu et al., 2011). Although this method is based on the PM equation, PM-MOD does neither require wind speed nor soil moisture data for the parameterization of aerodynamic and surface resistance. Further details about PM-MOD can be found in Mu et al. (2011) and Running et al. (2019). Note that some updates have been implemented in PM-MOD since Mu et al. (2011), which can be found in Running et al. (2019). These were also considered here in the implementation of PM-MOD.

2.3.4. PM-VI

This model relies upon the hypothesis that ET is mostly controlled by specific dominant processes, such as transpiration and photosynthesis, hence a good correlation between such processes and ET is necessary for good model performance (Nagler et al., 2007). There are several formulations to estimate ET from VIs (Nagler et al., 2005, 2009). In this study, we selected the algorithm proposed by Nagler et al. (2013),
which estimates \( ET \) using the reference crop evapotranspiration, \( ETo \), from the FAO-56 Penman-Monteith (PM) equation (Allen et al., 1998), and a crop coefficient, \( Kc_{VI} \), derived from a vegetation index. \( Kc_{VI} \) can be calculated in different ways (Nagler et al., 2005, 2013). Following Nouri et al. (2016) and Oliveira et al. (2015), \( Kc_{VI} \) was calculated as:

\[
Kc_{VI} = a(1 - e^{-b\cdot\text{EVI}}) - c
\]  

(1)

where \( a, b, \) and \( c \) are fitted coefficients. We used a parameter optimization tool based on a genetic algorithm to optimize the coefficients to estimate \( ET \) values close to the measured ones (Oliveira et al., 2015). The fitting procedure minimizes the objective function (OF) given by the sum of squared differences between tower-based \( ET \) \( (ET_{obs}) \) and \( ET \) estimates from the models \( (ET_{sim}) \) at time :

\[
OF = \sum_{i=1}^{n} \left[ET_{obs}(i) - ET_{sim}(i)\right]^2
\]  

(2)

This model, herein referred to as PM-VI, has frequently been employed to estimate \( ET \) at local and regional scales (Jarchow et al., 2017; Nouri et al., 2016; Oliveira et al., 2015). Although obtaining \( ETo \) requires a considerable amount of meteorological variables, the PM-VI implementation is easier and has a lower computational cost compared to other models. Unlike the three other models, PM-VI requires the calibration of the fitting coefficients, which can be a major issue for regions where \( ET \) and VI are poorly correlated or when correlations change over time (Chong et al., 1993). To calibrate the fitting coefficients, we randomly selected 20% of the available data at each site and used the remaining 80% to validate the model.

2.4. Quantifying Model Reliability

The model predictive skill was visually evaluated with scatter plots of measured versus modeled \( ET \), as well as through the coefficient of determination \( (R^2) \), root mean square error \( (RMSE) \), percent bias \( (PBIAS) \), concordance correlation coefficient \( (\rho) \), slope \( (m) \), and intercept \( (b) \) of the linear regression. The data used in the analysis were filtered for rainy days \( (P > 0.5 \text{mm}) \). Our analysis proceeded from a general (no distinction among sites) to a site-by-site and group level analysis, that is, per biome, climate, or land use. The number of flux towers assigned to each subgroup (i.e., the different biomes, climates, and land use classes) varied, and so did the record length per subgroup. To account for the different sizes, the following sampling procedure was performed, in which we computed the variability of each performance metrics for each group analysis (i.e., across its different subgroups):

(a) A sample size \( N \) was defined as half of the record length of the shortest subgroup, among all models; (b) for each model, samples of length \( N \) were taken from within each subgroup, and the performance metrics were computed; (c) This procedure was repeated 1,000 times, yielding a mean and standard deviation \( (SD) \) of the metrics at each subgroup, per model. The resulting SD are likely to be influenced by the choice of \( N \), and other rationale for its choice could have been made. In this way, the confidence bands reported here are to be seen as measures of relative variability, that is, the variability between the models, and not as absolute uncertainty bounds for each of them. To establish a relationship between model predictive skill and water availability at individual tower sites, we obtained the aridity index \( (AI = P/ETo) \) from the global dataset provided by Trabucco and Zomer (2019). For many tower sites, the available meteorological data (even from nearby meteorological stations) were not sufficient to provide a reliable \( AI \); hence the choice for a global dataset.

3. Results

3.1. ET Partitioning

Partitioning of \( ET \) among the three components \( (E_{tran}, E_{int}, \) and \( E_{soil}) \) exhibited more variation for the PT-JPL and PM-MOD models. On average, \( E_{tran} \) accounted for 60% (PT-JPL) and 56% (PM-MOD) of \( ET \) but, across sites, it presented a smaller range (30%-85%) for PT-JPL than for PM-MOD (20%-90%) (Figure 2, Table 1). GLEAM \( E_{tran} \) accounted for 82% of \( ET \) on average, varying between 60% and 95% across sites. Average interception across sites reached 9% (GLEAM), 13% (PT-JPL), and 24% (PM-MOD) of total \( ET \). \( E_{soil} \) fractions range were similar for GLEAM and PT-JPL \( (SD \approx 9\%) \), whereas PM-MOD \( E_{soil} \) varied more among sites \( (SD = 18\%) \). \( E_{int} \) was often correlated with \( LAI \), especially for the GLEAM estimates.
PT-JPL soilEE estimates exceeded the other models, particularly for sites with low \( E \) values (e.g., ESEC, CST, and USR).

3.2. Overall Model Skills

Since each model requires a different input dataset (Table S3 in Supporting Information S1), the data available to run and validate each model varied. GLEAM and PT-JPL provided a significantly greater number of daily outputs: 7301 (GLEAM), 7277 (PT-JPL), 5905 (PM-MOD), and 6638 (PM-VI). The complete dataset was used to produce scatter plots of \( E \) records and model simulations for each location (See Figures S4–S7 in Supporting Information S1). To allow a fair analysis, the results shown in the main text were obtained using data from days that were common across models, resulting in 4,718 data points.

To illustrate the relative contribution of each site to the scatter plots in Figure 3, we display the regression lines (light gray lines) between model and tower-based \( E \) for each tower site, and the mean metrics across individual sites. In general, \( E \) was reasonably predicted by all models, as suggested by the relatively low spread of most points in the scatter plots, many regression lines close to the 1:1 line, mean determination coefficient, \( R^2 \), mean concordance correlation coefficient, \( \rho \), mostly above 0.65, and mean root mean square error (RMSE) below 1 mm d\(^{-1}\) (Figure 3). Nevertheless there is some spread for a few sites, for example, in the PT-JPL scatter plot that displays a few sites with large bias despite strong overall correlation and \( \rho \).
The models slightly overestimate ET as suggested by higher density of points below the 1:1 line, except for GLEAM, which slightly underestimates. Correlations were similar between GLEAM and PT-JPL, with an average value of ~0.65 and the highest values at individual sites reaching close to 0.9, as indicated by the standard deviations (0.19 and 0.18, respectively). From Figures 3 and 4, it becomes evident that, despite the relatively lower spread of points for PM-VI, compared to the other models, this model (PM-VI) presented a less consistent performance across towers, as suggested by the contrasting slopes presented by the regression lines in that plot (e.g., reversed trend line at K77); hence the lower average determination coefficient ($R^2$) and $\rho$. Such a contrasting aspect of the PM-VI model is also noted by the fact that a wide range of $R^2$ was found despite the similarity between mean simulated and observed ET (Figure 4).

### 3.3. Model Skills Per Biome, Land Use, and Climate

Figure 5 presents $\rho$, RMSE, $\text{PBIAS}$, and $R^2$ for each model across six biomes, eight land use types, and seven climate classes in South America. Error bars are shown for all metrics and they represent the standard deviation resulting from the resampling procedure outlined in Section 2.4. Note that the analysis about the FGS and TBMF biomes are based on one and three towers, respectively. For most biomes, RMSE and $R^2$ did not significantly diverge. In general, TSGSS showed the best overall metrics for all models, while PM-VI in FGS (NPW site) presented the poorest ($\rho<0.5$, RMSE>1.5 mm d$^{-1}$, and $R^2<0.25$). Model performance

---

**Table 1**

**Comparison of Evaporation Fractions for Several Land Uses Between This Study and Field-Based Estimates**

<table>
<thead>
<tr>
<th>LULC</th>
<th>FE</th>
<th>GLEAM</th>
<th>PT-JPL</th>
<th>PM-MOD</th>
<th>FE</th>
<th>GLEAM</th>
<th>PT-JPL</th>
<th>PM-MOD</th>
<th>FE</th>
<th>GLEAM</th>
<th>PT-JPL</th>
<th>PM-MOD</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNF*</td>
<td>50–81</td>
<td>84–94</td>
<td>64–84</td>
<td>78–90</td>
<td>NA</td>
<td>8–14</td>
<td>14–24</td>
<td>8–18</td>
<td>10</td>
<td>1–2</td>
<td>2–5</td>
<td>2–4</td>
<td>de Queiroz et al. (2020); Gaj et al. (2016); Sun et al. (2019)</td>
</tr>
<tr>
<td>CROP*</td>
<td>NA</td>
<td>93</td>
<td>63</td>
<td>70</td>
<td>20–4</td>
<td>6</td>
<td>31</td>
<td>21</td>
<td>10</td>
<td>1</td>
<td>6</td>
<td>9</td>
<td>Cabral et al. (2012); Denmead et al. (1997)</td>
</tr>
<tr>
<td>CROP*</td>
<td>85</td>
<td>88</td>
<td>55</td>
<td>69</td>
<td>NA</td>
<td>3</td>
<td>20</td>
<td>4</td>
<td>13</td>
<td>9</td>
<td>25</td>
<td>27</td>
<td>Cabral et al. (2010)</td>
</tr>
<tr>
<td>WS*</td>
<td>NA</td>
<td>86</td>
<td>76</td>
<td>78</td>
<td>NA</td>
<td>2</td>
<td>17</td>
<td>5</td>
<td>8</td>
<td>13</td>
<td>7</td>
<td>17</td>
<td>Cabral et al. (2015)</td>
</tr>
</tbody>
</table>

**Note.** FE = field estimates. Land covers that present field data from the same modeling sites or same geographical region are indicated with “*”. 

The models slightly overestimate ET as suggested by higher density of points below the 1:1 line, except for GLEAM, which slightly underestimates. Correlations were similar between GLEAM and PT-JPL, with an average value of ~0.65 and the highest values at individual sites reaching close to 0.9, as indicated by the standard deviations (0.19 and 0.18, respectively). From Figures 3 and 4, it becomes evident that, despite the relatively lower spread of points for PM-VI, compared to the other models, this model (PM-VI) presented a less consistent performance across towers, as suggested by the contrasting slopes presented by the regression lines in that plot (e.g., reversed trend line at K77); hence the lower average determination coefficient ($R^2$) and $\rho$. Such a contrasting aspect of the PM-VI model is also noted by the fact that a wide range of $R^2$ was found despite the similarity between mean simulated and observed ET (Figure 4).
across towers within each biome did not vary much, as suggested by the relatively low range of the error bars for all metrics.

The central panels in Figure 5 provide evidence for the high variability of model predictive skills across different land uses (LU), which suggest that: (a) no model outperforms the others for all LU types and (b) each model has intrinsic and in some cases exclusive characteristic that makes it more suitable for certain LU. Only for croplands (CROP) we found similar metrics among models ($\rho \approx 0.8$, $0.8 < RMSE < 1.2$ mm d$^{-1}$, $20\% < PBIAS < 10\%$, $0.6 < R^2 < 0.8$). Conversely, for most LU, the metrics variation is remarkable (e.g., DBF: $0.4 < \rho < 0.9$, $-50\% < PBIAS < 10\%$, $0.25 < R^2 < 0.80$). On average, each model has the best skills across towers within each biome did not vary much, as suggested by the relatively low range of the error bars for all metrics.
for two LU; for example, ET prediction for GRA and DBF was best with PT-JPL ($\rho \approx 0.9$, $\text{RMSE} \approx 0.5$ mm d$^{-1}$, $\text{PBIAS} \approx 0\%$, $R^2 > 0.75$) whereas PM-VI presented similar skills for estimation of ET for CROP and PW. Likewise, model skill is related to the climate type. The analysis of $\rho$ and $R^2$ over semi-arid regions (BSk and BSh) indicates a relatively poor skill of all models (except PM-MOD for BSh climate). This is in contrast to the overall good performance over more humid environments (e.g., Aw and Cwb). The greatest divergence among model performances was found for the Polar Tundra (Td) climate zone, for which PM-VI presented the highest $\rho$ and $R^2$ (both $>0.75$), lowest $\text{RMSE}$ ($\approx 0.5$ mm d$^{-1}$) and $\text{PBIAS} (<10\%)$.

### 3.4. Individual Sites

In this section, we explore the model performance at individual towers. Model skills for all individual sites are depicted in Figure 6. Sites with $N < 30$ (CAX and MCR) are not discussed here but are considered in the scatter plots shown in the SM (Figures S4–S7 in Supporting Information S1). To facilitate the comparison of our results with previous analyses using the same models, only three statistics are shown in Figure 6: $\text{RMSE}$, $\text{PBIAS}$, and $R^2$. Other metrics are displayed in the scatter plots in Figures S4–S7 in the Supporting Information S1. In Figure 6, the metrics for the various towers are displayed in order of increasing arid-
ity (varying from $-\infty$ to 0, left to right), as suggested by the AI as described in Section 2.4. In general, there is a good agreement between the PM-based models in terms of RMSE and PBIAS.

Figure 5. Model performance per biome, land use, and climate. The error bars represent the standard deviation of the metrics within each class. Biome types: Tropical and Subtropical Moist Broadleaf Forests (TSMBF); Tropical and Subtropical Dry Broadleaf Forests (TSDBF); Temperate Grasslands, Savannas, and Shrublands (TGSS); Temperate Broadleaf and Mixed Forests (TBMF); Tropical and Subtropical Grasslands, Savannas, and Shrublands (TSGSS); Flooded Grasslands and Savannas (FGS); Land use types: Cropland (CROP); Woodland Savanna (WS); Deciduous Needleleaf Forest (DNF); Evergreen Broadleaf Forest (EBF); Grasslands (GRA); Mixed Forest (MF); Permanent Wetland (PW); Deciduous Broadleaf Forest (DBF). Climate Zones: Tropical monsoon (Am); Tropical savanna (Aw); Hot semi-arid (BSh); Cold semi-arid (BSk); Temperate oceanic (Cfb); Dry-winter subtropical highland (Cwb) and; Polar Tundra (Td). The sample size of each class is indicated in the x-axis.

In terms of individual metrics, RMSE values varied between $-0.5$ and $-1.5$ mm d$^{-1}$ for all models, with RMSE $<1$ mm d$^{-1}$ for most sites. The boxplots show that RMSE variation is similar among models, except...
for PT-JPL, which presents the lowest RMSE (e.g., K67). Figure 6 shows that PBIAS for PM-VI varies around zero across sites, which is expected given the model requires calibration with local data. However, based on $R^2$, it is apparent that this model's skill is quite limited for $AI \gtrsim 1.2$ and $AI \lesssim 0.5$. In general, the PT-based models showed larger biases, with PT-JPL and GLEAM consistently overestimating and underestimating $ET$, respectively. In terms of $R^2$, the PT-models ranked better than the PM-models for more than ~50% of the towers.
4. Discussion

4.1. General Implications

We conducted the first multi-remote sensing ET model analysis in South America (SA) using a common set of forcing and validation data located on flux tower sites across a diverse range of land covers, climates, and biomes. Forcing data include both in situ (e.g., temperature and net radiation) and remote sensing data, mainly related to vegetation (e.g., \( \text{LAI} \) and \( \text{ETVI} \)). To evaluate the models, energy-balance-derived \( \text{ET} \) (\( \text{ET}_{\text{EB}} \)) was used as observation, instead of eddy covariance \( \text{ET} \) (\( \text{ET}_{\text{EC}} \)). Given the benefits and drawbacks of using either \( \text{ET}_{\text{EB}} \) or \( \text{ET}_{\text{EC}} \), we compared both measures to verify whether such choices would lead to different results. As shown in the SM, for the great majority of tower sites, \( \text{ET}_{\text{EC}} \) and \( \text{ET}_{\text{EB}} \) are similar (Figure S10 in Supporting Information S1) and model statistics (\( R^2 \), \( \text{RMSE} \), and \( \rho \)) remained the same regardless of the ET approach or indicate a better performance when \( \text{ET}_{\text{EB}} \) was used (Figures S11–S13 in the Supporting Information S1). Many of the tower sites considered here are not yet available in flux network databases, including sites with land cover (deciduous needle-leaf forests, DNF), a biome (FGS), and two climate types (polar tundra and hot semi-arid) that have not been previously assessed in other regional studies on the performance of satellite-based ET models. Moreover, some classes included here were considered for validation of individual models only (e.g., semi-arid and tropical climate types, and TSDBF biome).

The fulfillment of such gaps (i.e., model evaluation across uncharted regions) is an important step because it allows a multitude of applications and studies relying on large scale ET mapping, such as: drought monitoring (Anderson et al., 2011, 2016), agricultural water management (Anderson et al., 2012), and diagnosis of climate change (Mao et al., 2015). The current ability to map ET remotely at various spatial and temporal scales, could only be evaluated thanks to the vast number of eddy covariance towers available in continental and global flux networks. As shown in this study, a thorough assessment of RBSET models based solely on data from such networks would be challenging or insufficient for some regions or continents; hence the relevance of this study. Our analysis provides essential information to identify model strengths and limitations across SA, allowing the users to identify which model is more suitable for them. Knowing under what circumstances (e.g., land use or climate) each model is more reliable is necessary to address remaining research and applied science gaps relative to ET at local, regional, and global scales (Fisher et al., 2017). Despite the value of tower-based ET across SA, many of those questions persist due to our limited observational capabilities. According to Fisher et al. (2017), the way to begin answering those questions is producing high quality ET estimates, which includes acquiring accurate ET information at high temporal and spatial resolution with large spatial coverage for a sufficient long period.

4.2. Model Performance, Sources of Errors, and ET Partitioning

Generally, model predictive skill over SA resembles what has been reported for other continents, including satisfactory values of coefficient of determination (\( R^2 > 0.6 \)) of the models (except PM-VI) for most validation sites, and consistently better results for the GLEAM and PT-JPL models, with \( \text{RMSE} \) ranging from \( \sim 0.5–1.5 \text{ mm d}^{-1} \) (McCabe et al., 2016). Also, in accordance with previous analysis, GLEAM and PT-JPL presented somewhat higher \( \text{RMSE} \) than PM-MOD but no clear evidence indicates decreasing performance with increasing aridity, as reported by McCabe et al. (2016); and Michel et al. (2016). Nonetheless, the general analysis (Section 3.2) indicates that all models can be used reliably over most of the environmental conditions in SA covered in our study. The analysis across towers and groups (i.e., biome, land use type, and climate, Section 3.3, Figure 5) identified considerable differences in terms of model skill.

Our results agree with previous studies from (Ershadi et al., 2014; McCabe et al., 2016; Michel et al., 2016; Miralles et al., 2016) who applied PM-MOD, GLEAM (except Ershadi et al., 2014), and PT-JPL to sites located in Africa, Asia, Australia, Europe, and Middle East and reported that PM-MOD showed, for most sites, lower correlations with measured \( \text{ET} \) compared to GLEAM and PT-JPL. Unlike previous analysis, our study agrees with Michel et al. (2016) in the sense that model skill seems to be unrelated to land cover. Michel et al. (2016) also reported a wide variation of \( R^2 \) (0.2–0.8) and \( \text{RMSE} \) (0.8–2 mm d\(^{-1}\)), for different sites under mixed forests. Conversely, contrasting results between our results and previous studies were found
for woodland savanna. While we found $0.5 < R^2 < 0.8$ and $0.7 < RMSE < 1.5 \text{ mm d}^{-1}$, Michel et al. (2016) reported $R^2 < 0.2$ and $1 < RMSE < 3 \text{ mm d}^{-1}$.

Overall, our group-wise analysis based on climate agrees with previous studies. For example, the poor model skill found here for the cold semi-arid (Bsk) climate (0.1 $< R^2 < 0.5$) resembles that found by McCabe et al. (2016) and Michel et al. (2016) for several sites in the United States. While aridity could have played a role here, it could also be caused by the fact that semi-arid sites are covered with sparse canopies. Such canopies present challenges when it comes to the description of aerodynamic transfer for example and radiation partitioning (see e.g., Verhoef & Allen, 2000). Our findings also show a poor to moderate model skill for $ET$ predictions for sites located in the Cfb climate zone, with PM-MOD having the worst performance. Conversely, PM-MOD presented the best predictive skill for the BSh climate, according to most metrics.

Besides the three RSBET models commonly assessed (GLEAM, PT-JPL, and PM-MOD), our analysis included the PM-VI model, which has been validated mostly for cropland or riparian ecosystems (Jarchow et al., 2017; Nagler et al., 2005, 2009, 2013). Here, we tested PM-VI for a much wider variety of biomes, climates, and land uses, and found a poor predictive skill for several sites with $AI > 1.2$ (e.g., K67, K77, and K83) or $AI < 0.5$ (e.g., CAA and SLU), even though the model accounts for a site-specific calibration. Considering the good results obtained for ~50% of the towers and the fact that, compared to the other models, PM-VI has a much simpler implementation, this model does have potential as long as sufficient data are available for calibration or, at least, validation. However, the need for local calibration is a hurdle for its implementation for most regions that are unsampled; therefore future studies are necessary to investigate which factors are most relevant in the determination of the model fitting coefficients, and to provide distributed reference values for its coefficients (e.g., based on land use dynamics).

We were able to identify a number of probable causes for poor model performance at individual sites, including (i) patch-scale heterogeneities; (ii) “mixed pixels,” that is, mixed response of different vegetation types within a pixel; (iii) time-lag between $ET_{\text{sh}}$ and $EVI$; (iv) low correlation between $ET$ and vegetation indices; and (v) model sensitivity to individual inputs (see Section 3.0 in the Supporting Information S1 for more details). Although we did not verify this in our study, we did not dismiss the possibility that known uncertainties in the estimation of site-specific vegetation characteristics (e.g., $f_{\text{PAR}}$ and leaf stomatal conductance in the PM-MOD; Ershadi et al., 2014); are further causes of lower model performance.

In our study, we used soil heat flux ($G$), which is generally measured below ground (usually at 5–20 cm deep) using soil heat flux plates. It could be argued that not correcting $G$ for the heat storage between the plate and the soil surface could lead to sub-optimal estimates of $ET$ when $LE$ is calculated as the residual of the energy balance, especially for towers where the soil is bare or covered by sparse vegetation, where $G$ can be relatively large. This, in turn, could lead to the conclusion that the models are performing worse than is actually the case. Although desirable, correcting $G$ for heat storage is rarely possible due to data unavailability (few sites only measure soil moisture and temperature, which are required to estimate soil heat capacity and heat storage using the calorimetric method). Moreover, at daily scales and for most sites, $G$ is either negligible in SA (summer or winter, when the amount of heat stored during the day roughly equals that lost during the night) or represents a minor portion only (spring and autumn) of the energy balance. As detailed and discussed in Section 3.0 and Figure S7 in Supporting Information S1, it is highly unlikely that neglecting such corrections will have affected the results.

There are, however, some issues worth mentioning here. Cause (v), for instance, is a major issue for PM-VI, as expected because the model is highly dependent on VI dynamics (see Section 2.4) (Nagler et al., 2005). Regarding cause, the superior performance of the PT models over PM-MOD at most sites is probably linked to uncertainties resulting from the estimation of aerodynamic resistance (Ershadi et al., 2014). In PM-MOD, the aerodynamic and surface resistances of each $ET$ component (soil, interception, and transpiration) are parametrized based on biome-specific values of leaf-scale boundary layer conductance, for example (Mu et al., 2011). Compared to the previous version of PM-MOD (Mu et al., 2007), this new approach resulted in a perceivable improvement only for cropland and deciduous broadleaf forest flux tower sites, whereas for other land uses no meaningful change was reported (Ershadi et al., 2015). Conversely, PT models are highly dependent on $Rn$ (causes iv and v); hence they often fail in dry environments (see metrics for $AI < 0.6$ in Figure 6), where $ET$ seasonality is dictated by $P$ more than radiation, or in regions with low $Rn$ (e.g., TF2).
Poor model responses at K77 (cropland, Figure S9 in Supporting Information S1) were attributed to causes (i) and (ii), as remnants of forest and shrubs were identified within the tower footprint and within MODIS pixel. VI products with higher resolution than MODIS exist and have been used to estimate ET (Aragon et al., 2018; Fisher et al., 2020); thus offering a possible solution for causes (i) and (ii). Time lag between ET and EVI (cause iii) was identified at EUC, where EVI followed the decline of ET after ~1–2 months.

Regardless of all those potential causes for poor model response, it is also important to consider to role of the core formulation upon which those RBSET models are based, that is, Penman-Monteith (PM) and Priestley and Taylor (PT) equations. A major problem of the PM equation refers to the linearization of the Clausius-Clapeyron relation, which has been addressed in a new version of that equation (McColl, 2020). The PT equation, in turn, implicitly assumes Rs and surface temperature (T) to be independent of evaporation. In reality, as shown by Yang and Roderick (2019), Rs not only decreases with increasing T (due to an increase of outgoing longwave radiation) but also a greater fraction of Rs becomes available for evaporation. Some of the deviations from the observations found in our analysis may happen due to such inconsistencies or simplifications. Here, we provide evidence to consider revisiting not just parameter values but the governing equations themselves and, ultimately, evaluate the benefits of such potential improvements in RBSET models.

Remote sensing based ET partitioning is expected to present some divergences from ground based measurements. This is the case especially for ET, because of the difficulty in obtaining remote sensing information on soil characteristics that drive ET. Such as soil moisture and temperature (Talsma, Good, Jimenez, et al., 2018; Talsma, Good, Miralles, et al., 2018), in particular at high vegetation cover fractions. Globally, transpiration has been reported to account for 57%–90% of global ET, based on in situ data and model outputs (Jasechko et al., 2013; Paschalis et al., 2018; Wei et al., 2017). Although these are global estimates, we expected ET to be the largest ET component also in SA due to its prevailing tropical climate and corresponding vegetation types. Our results show that this was indeed the case for GLEAM with an ET / ET ratio of ~80%, and for PT-JPL and PM-MOD with values of 57% and 60%, respectively. Nonetheless, based on our findings, model predictive skill in estimating total ET is not necessarily associated with its ability to partition ET accurately.

Concomitantly, inconsistencies in ET partitioning do not necessarily translate into inaccurate model estimates of total ET: this depends on the modelling approach. On the one hand, if total ET results from the sum of ET components independently, then an under- or overestimation of ET components can reduce the overall model skill, or reasonable ET estimates can be achieved as the consequence of an occasional compensation of errors in ET, ET, and ET. On the other hand, if the ET partitioning is derived from the estimate of a proxy value for total ET, such as available energy flux (as in PM-MOD and PT-JPL), the ET partitioning is unlikely to influence the total ET estimates. Moreover, ET partitioning may be sensitive to certain model inputs. For example, contrasting ET fractions were estimated by PM-MOD for similar rain forest sites, i.e., K67 and K83 (Figure 2). The reason PM-MOD is returning that difference is because RH was estimated from actual (e) and saturation vapor pressure (e) data, as RH data is not available in the K83 dataset. As a result, the difference between e-derived daytime and nighttime RH for K83 is greater than that for K67. In terms of daily averages, e-derived RH and measured RH are quite similar, which explains why the fractions for GLEAM and PT-JPL, at those two towers, are similar. Still, good estimates of ET components are important to differentiate the roles of vegetation and soil, that is, how they contribute to vertical soil water fluxes and changes in profile soil water content. Reliable knowledge of the distribution between ET and ET is also important when this information is used in hydrological models to calculate other water balance components, such as runoff.

Ground-based ET partitioning data are generally not widely available. This also goes for most land cover types included in this study. We compared the models’ outputs with field experiment studies that measured one or more ET components either at the same sites as those used here or within the same region (Table 1). ET partitioning values derived from GLEAM seem to be more consistent with ground-based information available for tropical rain forests, croplands and grasslands than for wetlands, and mixed and deciduous needle-leaf forests (Table 1). This also applies to PT-JPL with its ET partitioning agreeing reasonably well with observations made for both tropical rain and dry forests. Note that PT-JPL (as well as PM-MOD) constrain ET based on f_wet. Hence, compared to GLEAM, transpiration will be lower under high RH in the model
but ET can be high due to water availability in the soil and intercepted rainfall. Nonetheless, the overall predictive skill of PT-JPL was satisfactory at such sites (Figure 6 and Figure S4 in Supporting Information S1). Regarding PM-MOD, the main inconsistency is the $E_{trans}/ET$ for tropical forests (Table 1). Despite the wide variability in $E_{trans}/ET$ among models, their overall predictive skill was satisfactory, that is, not associated with their capability to correctly estimate each ET component individually (see Supporting Information S1 for further discussion). No model was able to consistently capture the ET partitioning across all sites correctly, which is expected given the uncertainty of each ET component and the climate and land-cover variability in SA. However, the joint estimates of all models covered totally or partially all field-derived evidence on ET partitioning. This suggests that continental ET estimates for understudied regions, such as the SA, would benefit from merging ET outputs from models that are based on different methods (Paca et al., 2019).

Despite our efforts to gather as much tower data as possible, with the goal of having a common dataset for all models, we faced several limitations including: differences in lengths of observational time series across towers (up to 3 years), as well as lack in overlap of these time series; uneven distribution of towers across groups (e.g., biomes); and, finally, South American geographical features that were not considered in this study (e.g., MGS biome or desert climate type, BWk). Thus, it was not possible to assess, for all towers, model responses during all seasons. Nonetheless, the fact that our dataset encompasses a wide variety of climates enabled us to evaluate, to a reasonable extent, model responses for contrasting seasons and fill in the gaps flagged by the literature, such as the absence, in a similar analysis, of towers in the tropical climate zone pointed out by McCabe et al. (2016).

5. Conclusion

Our results show that, in general, ET can be reasonably well predicted by all four models, despite an overall tendency of overestimation by PT-JPL and PM-MOD, and underestimation by GLEAM. Contrasting with results from other continents, we found no clear evidence linking model predictive skill with aridity. Our analysis emphasizes the need of improving model ET partitioning, although the link between flawed ET partitioning and poor model skill is not evident based on our results. Having reliable ET partitioning coefficients as part of the FLUXNET-type datasets would be valuable in this respect, but unfortunately such data are difficult to obtain, as they require labor-intensive and costly methods (such as sapflow gauges and lysimeters), that also present problems with regards to upscaling from plot to field-scale.

Correlations are consistently higher for GLEAM and PT-JPL, with $R^2 > 0.5$ for most sites, whereas PM-MOD and PM-VI presented better performances in terms of PBIAS ($-10 < PBIAS < 10\%$ for most sites). As for PM-VI, the low PBIAS is expected, given the model requires calibration with local data.

The model skill for the various models seems to be unrelated to land cover type as we found a wide variability of metric values within the same class and across models. Conversely, a relatively lower performance was observed for most models in semi-arid regions (e.g., BSk climate type), compared to an overall good performance for more humid environments (e.g., Aw climate type). Except for the FGS biome, we found that skill across models was mostly similar within the same biome.

Despite the relatively high number of towers (compared to previous global analyses that used a similar amount of sites), gathering a balanced amount of data and uniform distribution of towers across different biomes and climate zones across the whole continent was challenging. Thus, there is a need to expand the flux tower network in South America as well as the formation of bilateral collaborations for future contributions. Previous studies (McCabe et al., 2016; Michel et al., 2016) have expressed the need to extend the evaluation of RSBET models to uncharted biomes and climate conditions. Our analysis fills this gap by assessing the reliability of four RSBET models over South America. We provide benchmarking metrics that can serve the improvement of ET models for improved capturing of ET over this continent.

**Data Availability Statement**

The data used in this study is available at https://doi.org/10.5281/zenodo.5549321.


Pastorello, G., Trotta, C., Canfora, E., Chi, H., Christiansen, D., Cheah, Y.-W., et al. (2020). The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. Scientific Data, 7(1), 225. (Number: 1 Publisher: Nature Publishing Group) https://doi.org/10.1038/s41597-020-0534-3


