Journal of Hydrology 645 (2024) 132209

Contents lists available at ScienceDirect

Journal of Hydrology

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

Evaluating a multi-step collocation approach for an ensemble climatological dataset of actual evapotranspiration over Italy

C. Cammalleri [a,*], M.C. Anderson [b], C. Corbari [a], Y. Yang [c], C.R. Hain [d], P. Salamon [e], M. Mancini [a]

[a] Politecnico di Milano, Dipartimento di Ingegneria Civile e Ambientale (DICA), Milan 20133, Italy
[b] United States Department of Agriculture, Agricultural Research Service, Hydrology and Remote Sensing Laboratory, Beltsville, MD, USA
[c] School of Integrative Plant Science, Cornell University, Ithaca, NY, USA
[d] Marshall Space Flight Center, Earth Science Branch, NASA, Huntsville, AL, USA
[e] European Commission, Joint Research Centre (JRC), Ispra 210327, Italy

ARTICLE INFO

Keywords:
ET
Water balance
Surface energy balance
Multi-model ensemble

ABSTRACT

Accurate estimations of actual evapotranspiration (ET) are key in a variety of water balance applications, but divergent results can be obtained due to the large range of available methodologies. The use of an ensemble approach is a suitable alternative, as it summarizes multiple sources in an optimized strategy. In this study, an expert-based multi-step collocation (MC) approach is tested to merge six ET datasets, with the aim of reconstructing a spatiotemporally-consistent monthly dataset for Italy in the climatological period 1991–2020. The merged products are: three water balance datasets (BIG BANG, LSA SAF, and LISFLOOD), two residual surface energy balance model datasets (SSEBop, and ALEXI), and the MODIS standard ET product. The merged product is analyzed for spatio-temporal consistency and evaluated using flux observations from 11 sites. On average, the merged product has higher accuracy (mean absolute difference = 0.47 ± 0.17 mm/d, relative difference = 27.9 ± 7.5 %) than any single base dataset, and it is characterized by limited bias (mean bias error = -0.17 ± 0.26 mm/d), high correlation ($r = 0.83 \pm 0.10$), and more uniform performance across sites. The merged ET dataset is accompanied by an estimation of the ensemble spread, which highlights large differences in ET estimates in some areas and periods characterized by severe water stress, such as in southern Italy during the summer. This large spread seems to be mostly driven by systematic differences among datasets, which affect the estimation of the reference climatology, suggesting how inter-model spread can have a defining role in further improving the merging strategies.

1. Introduction

Actual evapotranspiration (ET) is a crucial variable in the hydrological balance, representing the main loss of water from the land surface. Reliable estimates of ET play a major role in quantifying the water stress in the agricultural sector (Wanniarachchi and Sarukkalige, 2022), in informing water management of the water losses at river basin level (McKinney et al., 1999), and in providing dynamic information on boundary conditions fostering drought events (Wood et al., 2015).

ET measurements are generally deployed using complex indirect techniques, which limit both data availability and overall accuracy (Allen et al., 2011). Globally, ET observation sites are sparsely available, continuous long records are limited, and gaps in time series are frequent. For these reasons, simulation approaches are usually implemented for large scale ET estimates, including land surface models, hydrological balance, and remote sensing-based methods (Courault et al., 2005; Wang and Dickinson, 2012).

Due to the large variability in ET, the spatial and temporal scales of different ET products play a major role in defining their range of applications. While field-scale agricultural studies may require high spatial (10^0-10^1 m) and temporal (sub-daily and daily) resolutions, water balance studies at basin or regional scale, as well as drought analyses, commonly rely on moderate spatial resolution (10^3 m) and monthly aggregation (e.g., Anderson et al., 2011; Di Giovanni et al., 2023; Rossi et al., 2023; Salvati et al., 2008). In this study, given the spatial scale of interest (the entire Italian territory), the focus is on moderate spatial

* Corresponding author.
E-mail address: carmelo.cammalleri@polimi.it (C. Cammalleri).

https://doi.org/10.1016/j.jhydrol.2024.132209
Received 29 November 2023; Received in revised form 16 July 2024; Accepted 5 October 2024
Available online 22 October 2024

resolution models at monthly time steps.

Among the datasets available over Italy at moderate spatial resolution, the *Istituto Superiore per la Protezione e Ricerca Ambientale* (ISPRA) recently developed a national GIS-based simplified water balance model (BIG BANG) (Braca et al., 2021). A comparison by the authors with the MODIS (Moderate Resolution Imaging Spectroradiometer) standard monthly product (MOD16) at national scale highlighted lower ET during the winter months and higher ET during the summer months, while the regions with the largest difference at annual scale are in south and north-east Italy. The MOD16 product (Mu et al., 2011) has been extensively evaluated over a large range of climatic and land use conditions within the US, reporting an average mean absolute error of about 24 % at the monthly timestep, with largest differences in grassland and woody plants. This error value is within the range of uncertainty of eddy covariance observations, but still significantly large. Other validation studies of the Land Surface Analysis Satellite Application Facility (LSA SAF) daily ET product reports a similar range of uncertainty, where the product target accuracy of 25 % is achieved in only 50 % of the validation sites (LSA SAF, 2023).

Previous studies have proven how ET modelling can be particularly challenging in the Mediterranean area due to the large range of water regimes, the high degree of spatial heterogeneity, and the marked difference between winter and summer conditions and coastal and inland areas (Rana and Katerji, 2000). Indeed, intercomparison exercises at global scale using coarse spatial resolution (10^4-10^5 m) models at monthly timesteps (i.e., Pan et al., 2020) have highlighted a large degree of uncertainty in model estimates, most notably over arid and semi-arid regions. Another example is the intercomparison study between LSA SAF and MOD16 products over Europe by Hu et al. (2015), which reported major biases over Italy and low correlation compared to central Europe.

This brief overview highlights how no single dataset can be considered clearly superior across a large, heterogeneous territory as Italy, and how the overall uncertainty of operational ET products is larger than an ideal target in many cases. This major limitation of single-model estimates has been addressed, over the most recent years, by using multi-source ensemble methods (Singh et al., 2019). In this framework, the triple collocation (TC) approach (Stoffelen, 1998) has seen widespread application in hydrology, mainly including studies on precipitation (e.g., Dong et al., 2020; Duan et al., 2021) and soil moisture (e.g., Gruber et al., 2016; Xie et al., 2022). Quadruple and quintuple collocations are also discussed in the scientific literature (Chen et al., 2021), as well as more general frameworks for the multiple collocation problem (Pan et al., 2015; Vogelzang and Stoffelen, 2022).

In spite of the extensive scientific literature on the topic, no unanimous consensus on the optimal strategy to perform a collocation analysis has yet to be attained. The most common TC applications include conversion of the real data to climatological (or even standardized) anomalies and rescaling to an arbitrary reference, with both procedures aiming at increasing the adherence of the raw datasets to the underlying assumptions of the collocation methods. McColl et al. (2014) argued that applying TC to standardized quantities is not strictly needed, whereas applications on soil moisture data suggest that anomalies may better satisfy the TC base hypotheses (Miralles et al., 2010; Draper et al., 2013). Similar discussions are held about the rescaling to a common reference system, with arguments against this practice reporting the possibility of biasing the error estimates (McColl et al., 2014).

Another point of discussion is related to the handling of mutual dependence in the model errors (Pan et al., 2015), with the absence of cross-correlation becoming more and more unrealistic with the increasing number of the analyzed datasets, and structural errors become unresolvable (Zwieback et al., 2012). Chen et al. (2021) suggested the possibility to test the zero-error cross-correlation assumption in TC for different dataset pairs using the quadruple collation, whereas Pan et al. (2015) recommend a procedure that incorporates expert knowledge in a multi-step collocation (MC) where similar datasets are preliminarily grouped according to a-priori knowledge.

As previously mentioned, limited applications of the different collocation procedures are available in the scientific literature for ET compared to soil moisture and precipitation. A recent example of such application is reported in Li et al. (2023), where both triple and quadruple collocations are tested over northern Europe. Park et al. (2023) also tested TC to merge both evaporation and transpiration reanalysis and satellite products over China. In the first study, the collocation procedure is applied to deviations from the climatology, and then two different reference datasets are used to reconstruct ET values, whereas in the second work the TC is applied directly to ET values.

A key difference in the ET applications compared to soil moisture is the relevancy of the actual magnitude of the modelled fluxes, which are difficult to properly reconstruct due to a lack of a reference dataset for rescaling. This issue differentiates ET also from precipitation, for which a ground-based gridded reference dataset is often available. The lack of a proper reference may become a major source of uncertainty in ET estimates when significant systematic biases are observed among datasets. In this context, a proper quantification of the spread of the merged product becomes even more important to provide an assessment of the inter-model agreement.

The main goal of this study is to produce a dataset of actual ET for the entire Italian territory covering the climatological reference period 1991–2020 (WMO, 2017) at moderate spatial resolution (1-km) and monthly time scale, with the additional goal to include a robust assessment of the spread of the actual ET best-guess estimates. This is achieved by combining six datasets retrieved from multiple sources and available for the Italian territory. These datasets are based on a variety of modelling schemes, and include: 1) the outputs of the BIG BANG simplified monthly water balance from ISPRA, 2) the estimates from the Meteosat-based land surface model from LSA SAF, 3) the MODIS satellite standard product MOD16, 4) the dataset based on the operation simplified surface energy balance (SSEBop), 5) the estimates from the ALEXI (Atmosphere-Land Exchange Inverse) two-source energy balance model, and 6) the ET simulations from the CEMS (Copernicus Emergency Management Service) LISFLOOD hydrological model.

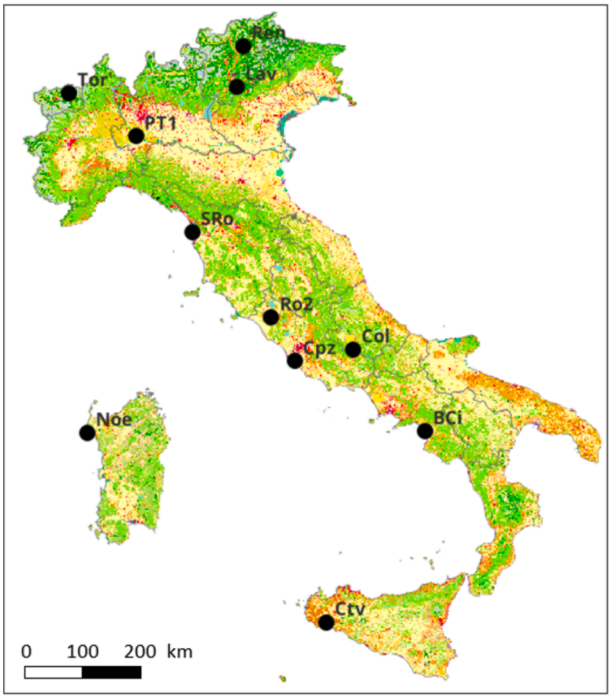
The combining strategy relies on the use of a MC-driven weighted ensemble procedure, with the objectives of: i) ensuring a temporally consistent merged dataset (albeit generated from base data sources with different temporal coverages), ii) preserving the interannual variability signals of the source datasets, as a key feature for the analysis of dry extremes (i.e., droughts), and iii) providing information on the spread of the estimates, as a proxy variable of the local agreement among the datasets in addition to the best-guess ensemble estimates. A particular focus will be given to the role of different components (climatology and anomalies) in the spread estimate, as a key factor in understanding potential further improvements.

2. Materials

2.1. Study area

This study focuses on the Italian national territory, encompassing the Alpine and sub-alpine areas in the north, the peninsular regions and the two major islands, Sicily and Sardinia (Fig. 1). The study area (about 300,000 km^2) covers different climatic regimes according to the Köppen classification, from humid continental and subtropical climates in the north, to a typical Mediterranean climate in the center and south. Subartic climate is also observed over the highest altitudes of the Alps (Pinna, 1970).

Relevant for the ET modelling are also the main land uses. According to the Corine land cover dataset (CLMS, 2021), about half of the Italian territory is classified as agricultural land, equally split between non-irrigated and permanently irrigated arable lands. Forests cover most of the remaining territory (about 40 %), with a predominance along mountain regions. Four macro-regions can be considered to concisely represent the main results of our analyses; these are based on the NUTS 1



ID	lat/lon	land cover	period
Tor	45.8,7.6	Grassland	2009-2014
Ren	46.6,11.4	Evergreen Needleleaf Forest	2002-2012
Lav	46.0,11.3	Evergreen Needleleaf Forest	2003-2014
PT1	45.2,9.1	Deciduous Broadleaf Forest	2002-2004
SRo	43.7,10.3	Evergreen Needleleaf Forest	1999-2007
Ro2	42.4,11.9	Deciduous Broadleaf Forest	2002-2012
Cpz	41.7,12.4	Evergreen Needleleaf Forest	1997-2008
Col	41.8,13.6	Deciduous Broadleaf Forest	1996-2014
BCi	40.5,15.0	cropland	2007-2012
Noe	40.6,8.1	shrubland	2005-2014
Ctv	37.6,12.8	perennial woody crop	2009-2011

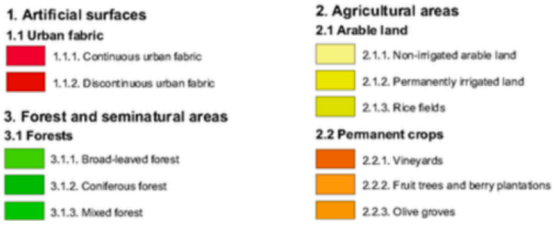


Fig. 1. Study area. The map reports the Corine land cover for the Italian territory, with location of the in-situ stations used in the analysis, as detailed in the table. The legend reports only the major land cover, with forest areas depicted in green shades, agricultural areas depicted in yellow to brown shades, and urban areas in red (see https://land.copernicus.eu/pan-european/corine-land-cover/clc2018 for the full legend of the classification).

administrative division, namely: North (including ITC and ITH), Central (ITI), South (ITF), and Insular (ITG) regions (Eurostat, 2022).

Monthly records of actual ET were collected from eddy covariance (EC) stations in the FluxNet network (https://fluxnet.org), and in particular 10 sites were selected from the FLUXNET2015 dataset (Pastorello et al., 2020) (see Fig. 1). This dataset is characterized by an improved quality control and post-processing, and includes data already aggregated at monthly scale and with balance closure applied (variable LE_CORR, latent heat with balance closure enforced). The quality of the observations used in the aggregation is summarized in a dedicated quality flag field (LE_F_MDS_QC, fraction of measured records in the aggregation period), and only data with a fraction greater than 0.9 were used. This filter, in addition to the use of balance-closed data (which require reliable observations of the other component of the energy balance as well), ensures the use of only high-quality data with a high level or replicability (since FluxNet standard values are used).

Stations were chosen to guarantee a minimum of 36 monthly observations across 4 consecutive years, with replicated stations or neighbor sites removed. To extend the spatial coverage to Sicily Island, the 'Ctv' station (Cammalleri et al., 2013), not part of FluxNet, was also included in the ground-based dataset. The data were post-processed using the same procedure adopted in FluxNet, and energy balance closure was forced by preserving the observed Bowen ratio, hence by redistributing the error to the turbulent fluxes proportionally to their magnitude (Twine et al., 2000). This closure method is the same adopted in FLUXNET2015 (Pastorello et al., 2020), and the most used procedure when balance closure must be ensured (as it is the assumption of all modelling frameworks).

The spatial homogeneity of the landscape surrounding the in-situ stations was also verified, given the relatively coarse spatial resolution of the modelled data (~ 10^3 m) compared to the typical EC footprint (~ 10^2 m). Nevertheless, even in highly heterogeneous landscapes, or for low resolution datasets, comparison against in-situ EC fluxes remains the most common practice in studies evaluating ET gridded estimates (e.g., Li et al., 2023; Yang et al., 2017; Zhu et al., 2022) given the lack of viable alternatives.

2.2. Modelled ET datasets

In this study, a preliminary analysis of the ET datasets covering the Italian territory with spatio-temporal characteristics compatible with the target goals was performed, in order to detect suitable candidates for the analysis. Six ET datasets (Table 1) were identified, all openly available, covering a large portion of the period of interest, characterized by similar moderate spatial resolution, and encompassing a range of methodological approaches. The datasets can be divided, based on the modelling approach, in three main groups: 1) water balance, 2) residual surface energy balance, and 3) Penman-Monteith based equation. A brief description of each dataset is provided in the next sub-sections, following this general schematization.

2.2.1. Water balance

The BIG BANG (*Bilancio Idrologico GIS based a scala nazionale su griglia regolare*) dataset (version 6.0, Braca et al., 2023) comprises all the main hydrological quantities (ET, soil moisture, runoff, snow, etc.), as simulated at the Italian national level using a simplified water balance model at monthly time scale and 1-km spatial resolution. The meteorological forcing of the model is obtained by interpolating in-situ observations from about 2,500 rain gauges included in the DBPLUVIOM

Table 1
Summary of the characteristics of the ET datasets used in this study.

Name	Resolution		Period	Reference
	Spatial	Temporal		
BIG BANG	1-km	Monthly	1991–2020	Braca et al. (2023)
LSA SAF	3.5-km	Daily	2011–2020	LSA SAF (2016)
LISFLOOD	1-km	6-hour	1991–2020	Van der Knijff et al. (2010)
SSEBop	1-km	Monthly	2003–2020	Senay et al. (2013)
ALEXI	5-km	Daily	2005–2020	Anderson et al. (1997)
MOD16	500-m	8-day	2001–2020	Running et al. (2017)

4.0 database. The soil hydraulic characteristics are derived from the data available in the European Soil Data Center (ESDAC) (Panagos et al., 2012). Due to the simplified nature of this method, the effect of irrigation supplies on the actual ET is not incorporated. ET estimates from BIG BANG have not been thoroughly evaluated against observations, but an intercomparison with the MOD16 product has shown negative/positive differences during the winter/summer months (Braca et al., 2021). At regional scale, differences up to 25 % with respect to MOD16 are observed for annual ET, with major differences over Calabria, Sardinia, Apulia (in the south), Veneto and Friuli Venezia Giulia (in the north-east). Since this product covers the Italian territory at the target spatial resolution, its grid is used as a reference for the resampling of all the other products.

The LSA SAF (Land Surface Analysis Satellite Application Facility) DMET satellite product provides estimates of actual ET based on a combination of Meteosat-generated meteorological forcing and surface variables (i.e., radiative forcing, albedo, leaf area index) and forecasts (air temperature, humidity and pressure, and wind speed) provided by ECMWF (the European Centre for Medium-range Weather Forecasts) Integrated Forecasting System (IFS). The adopted modelling scheme is a Soil-Vegetation-Atmosphere Transfer (SVAT) based on the H-TESSEL (revised Hydrology Tiled ECMWF Scheme for Surface Exchanges over Land) model (Balsamo et al., 2009). Landcover classes within each pixel are derived from the ECOCLIMAP-II dataset (Faroux et al., 2013), with a maximum of four tiles per pixel. The combined use of satellite and reanalysis data allows to easily integrate other input data from external sources in near-real time (Ghilain et al., 2011), including snow cover and soil moisture information from H-SAF (LSA SAF, 2016). The daily product is obtained by temporally integrating 30-min estimates, and it contains information on the number of reliable observations in each day (out of the 48 possible 30-min values). In this study, valid monthly maps are obtained if at least 20 valid days are available, with a valid day defined as a day with reliable data in at least 1/3 of the 30-min estimates. The spatial resolution of the product is roughly 3.5-km over southern Europe, and the data have been interpolated to a 1-km regular grid using a bilinear approach. A validation exercise conducted over the entire Meteosat disk reported a mean absolute error (MAE) of 0.9 mm/d over about 60 eddy covariance sites. In particular, over Italy the model performance varies substantially, yielding high correlations with observations in “Ren” and low correlation in “BCi” and “NOE” (see Fig. 1 for location), with a systematic tendency for negative biases in all the deciduous broadleaf forest sites (LSA SAF, 2023).

The LISFLOOD model is a physically-based hydrological balance model originally developed for flood monitoring (de Roo et al., 2000), which currently runs operationally at European scale as part of the European Flood Awareness System (EFAS) of the Copernicus Emergency Management Service (CEMS, https://emergency.copernicus.eu/) activities. In its latest operational version (v5.0), LISFLOOD simulates the soil water balance in three sub-layers for a set of sub-grid homogeneous land cover classes (forest, impervious surface, inland water, and agricultural areas) at 6-hourly time steps on a 30 arcsec regular grid. Meteorological forcing gridded maps are derived from an interpolation of local stations as provided in near-real time by a large set of partners (https://www.efas.eu/en/share-your-data-efas), with some of the same stations available in BIG BANG also used here. Irrigation is simulated only for the sub-grid classified as irrigated agricultural land, with two different sub-routines dedicated to crops and paddy-rice. The model is calibrated on river discharge over a large set of European basins (more than 1,900 gauges), covering some water basins in northern and central Italy but no areas in southern Italy (https://confluence.ecmwf.int/display/CEMS/EFAS+v5.0+-+Calibration+Methodology+and+Data). Analyses on soil moisture (Cammalleri et al., 2015) report a general good agreement with other water balance models, but also discrepancies over southern Italy during summer months. No specific validation of LISFLOOD has been performed on ET. In this study, daily ET values are obtained by summing all the sub-daily estimates, and the data were simply re-sampled on the BIG BANG grid using the nearest neighbor approach.

2.2.2. Residual surface energy balance

The SSEBop (operational Simplified Surface Energy Balance) model provides estimates of actual ET at global scale based on the resolution of a simplified surface energy balance approach (Senay et al., 2011). A peculiarity of this method is the use of pre-defined, seasonally dynamic, boundary conditions that are unique to each grid cell, analogous to the “hot/dry” and “cold/wet” reference points adopted in the SEBAL (Surface Energy Balance Algorithm for Land) and METRIC (Mapping EvapoTranspiration at high Resolution with Internalized Calibration) approaches (Bastiaanssen et al., 1998; Allen et al., 2007). The product used in this study (version 5.0) combines the ET fraction generated every 8 days from remotely sensed MODIS thermal imagery (incorporating the effects of water stress and irrigation on land-surface temperature; LST), with a climatological reference ET from gridded weather datasets merging data from the Global Data Assimilation System (GDAS) and the dataset from the International Water Management Institute (IWMI) (Senay et al., 2020). Comparison with 12 flux towers worldwide (no sites in Italy) showed a good capability to capture the seasonality (high correlation), but also bias ranging between 3 and 50 % with both negative and positive values. Monthly maps at the MODIS 1-km resolution are directly disseminated (https://earlywarning.usgs.gov/fews/product/460) by combining the 8-day estimates; monthly data are re-gridded on the BIG BANG grid with the nearest neighbor approach.

The ALEXI (Atmosphere-Land Exchange Inverse) model (Anderson et al., 1997; Anderson et al., 2007) simulates the surface energy fluxes under actual conditions by relating time-differential observations of morning LST rise to the time-integrated energy balance within the surface-atmospheric boundary layer system. The model solves the energy balance for vegetated and bare soil components of each pixel, separately, by partitioning the remotely observed LST between the soil and canopy contributions using the Two-Source Energy Balance (TSEB) land-surface representation of Norman et al. (1995). In the version of ALEXI used in this study, daily ET estimates at 5-km spatial resolution are derived from MODIS day/night thermal observations rescaled as described in Hain and Anderson (2017). MODIS data are also used to characterize surface albedo and the vegetation fractional coverage/leaf area index. Meteorological forcings, including solar radiation, wind speed, air temperature, and vapor pressure, are obtained from the Climate Forecast System Reanalysis (CFSR; Saha et al., 2010). Surface roughness used to compute aerodynamic coupling between land and atmosphere is defined based on the ISLSCP II (International Satellite Land-Surface Climatology Project, Initiative II) land-cover classification (DeFries and Hansen, 2010). The ALEXI model has been extensively tested over the US in combination with the DisALEXI approach, which spatially downscales ALEXI ET estimates to 30-m resolution using Landsat LST and vegetation cover fraction information (Anderson et al., 2004; Anderson et al., 2018). The most comprehensive U.S. based evaluations of both ALEXI/DisALEXI and SSEBop were conducted under the OpenET model intercomparison project (Volk et al., 2024), yielding monthly MAE of 25–28 mm/month (28–30 %) in croplands, with higher bias and error in forested landcovers. Comparison studies over Europe and Italy, however, are limited. Valid monthly estimates are obtained when at least 20 valid daily ET values are available, and successively interpolated to the BIG BANG grid with a bilinear approach.

2.2.3. Penman-Monteith based equation

The MODIS actual ET product (MOD16A2GF, collection 6.1, MOD16 from hereafter) is a year-end gap-filled 8-day composite dataset produced globally at 500-m spatial resolution. The algorithm used to generate this product is based on a revised version of the Penman-Monteith (Monteith, 1965) approach, in which details on albedo, biome type, vegetation cover fraction, and leaf area index are directly derived from other standard MODIS products, whereas the effects of water stress are simulated under the assumption that minimum air

temperature and vapor pressure deficit are the major constraints on stomatal conductance (Running et al., 2017). Meteorological forcing is obtained from the near-real time surface weather data generated by the Goddard Earth Observing System Model, Version 5 (GEOS-5), at GMAO/NASA, and includes solar radiation, wind speed, vapor pressure, and air temperature data. A validation study has been conducted by Mu et al. (2011) over the US, as well as over other regions in several analyses (e.g., Kim et al., 2012; Ramoelo et al., 2014). No comprehensive validation of this product over Italy is available in the scientific literature, but some local-scale studies (Autovino et al., 2016; Castelli, 2021) have shown a good accuracy in southern Italy but poor performance over the Alps. In this study, monthly ET maps are obtained as a weighted average of the 8-day estimates, with the weight corresponding to the overlapping days between the 8-day product and the monthly time window. The data are spatially upscaled to the BIG BANG 1-km regular grid as a simple average.

3. Methods

A methodology to produce a multi-model actual ET dataset, including information on inter-model spread, is described here. This methodology aims at achieving the three main objectives: 1) to ensure that the final product is characterized by an internal temporal consistency during the entire target period (1991–2020), even if the six base products cover different timeframes (see Table 1); 2) to optimize the ensemble strategy in order to produce a dataset that preserves not only the ET fluxes but also the interannual fluctuations at monthly scale (relevant for drought studies); and 3) to produce not only the best-guess estimates of monthly ET but also a quantification of the spread around the best-guess estimates, as a proxy metric of the overall consistency of the six models.

To this aim, starting from the common definition of deviations from the climatology, often also known as anomalies, we can express the monthly actual ET, ET_{my}, for a given grid cell as:

$$ET_{my} = \mu_m + z_{my} \tag{1}$$

where the subscripts m (1 – 12) and y (1991 – 2020) identify the month and the year, respectively, μ_m is the climatological monthly average, and z_{my} is the anomaly from the climatology (i.e., the interannual fluctuation as deviation from the climatology).

Analogously, the spread in the ET estimates, ε_{my}, can be expressed by the sum of the spread of the two sub-components:

$$\varepsilon_{my} = \varepsilon^{\mu}_{m} + \varepsilon^{z}_{my} \tag{2}$$

where ε^{μ}_{m} and ε^{z}_{my} are the spread of the climatology and anomaly estimates, respectively. Equation (2) assumes the absence of covariance between the two terms on the right-hand side.

In this approach, the climatology and anomaly components are modelled separately, which allows us to have a common baseline dataset for the entire period (internal temporal consistency, objective 1), and accounts for the possible systematic biases among datasets (included in the term ε^{μ}_{m}), while ensuring the robustness of the inter-annual fluctuations (objective 2) by directly combining them independently from the climatology. The procedures to estimate both the climatological quantities and the anomalies are described in the next sub-sections.

3.1. Estimation of the interannual fluctuations

The anomaly term in Eq. (1) can be modelled as a weighted average of the single dataset anomalies, z^{k}_{my}, as:

$$z_{my} = \sum\nolimits_{k=1}^{K} \omega^{k} z^{k}_{my} \tag{3}$$

where the superscript k identifies the base datasets (out of $K = 6$ models), and the weighting factors, ω^{k}, are computed following the multi-step collocation (MC) modelling framework introduce by Pan et al. (2015).

The method proposed by Pan et al. (2015) is based on a well-interpreted collocation technique under the framework of Pythagorean constraints in Hilbert space, which assumes that the variables are unbiased estimates with linear scaling error removed, as in the case of anomalies, and adopts as main quantity the root mean squared distance, d^2, between any two estimates in a Hilbert space. Other key hypotheses of the MC method, similar to those of the well-known triple collocation (TC) method, are the independency between the errors and the truth (error orthogonality) and the mutual independence of the errors among the datasets (zero error cross-covariance) (Gruber et al., 2016). Yilmaz and Crow (2014) highlighted how this latter term may be a source of bias in error estimates, and how the hypothesis of independence is more unrealistic when a large number of datasets is used.

In this framework, the errors d^2_{kt} for each k-th model compared to the unknown truth (subscript t) (K unknowns) are assessed starting from a set of C^2_k Pythagorean equations:

$$\left\{ d^2_{kt} + d^2_{jt} = d^2_{kj} \right\}_{\forall k \neq j} \tag{4}$$

where $C^2_K = K(K-1)/2$, and the subscript j represent the other non-k estimates. It is worth highlighting how Eq. (4) is a generalization of the TC problem, which can be applied to any number of datasets (even greater than 3). Indeed, the problem becomes over-constrained ($K < C^2_K$) if $K > 3$, and it is solved with a least squared solution that can be expressed in the form:

$$d^2_{kt} = \frac{1}{C^2_{K-1}} \sum\nolimits_{i \neq j} \varphi_{ij} d^2_{ij} \quad \text{and} \quad \varphi_{ij} = \begin{cases} K-2, & i = k \vee j = k \\ -1, & i \neq k \wedge j \neq k \end{cases} \tag{5}$$

An optimal set of weighting factors to be used in Eq. (3) can be estimated from the average errors derived from Eq. (5) as:

$$\left\{ \omega^{k} = d^{-2}_{kt} / \sum\nolimits_{k=1}^{K} d^{-2}_{kt} \right\}_{k=1,\cdots,K} \tag{6}$$

where this minimum squared error solution gives more weight to the estimates with smaller errors (Yilmaz et al., 2012).

As previously mentioned, Eqs. (5) and (6) can be directly applied to any number of datasets, which, in this case, would be the six datasets described in section 2.2 ($K = 6$). This approach would assume mutual independent errors among all the datasets (similar to the assumption in TC). In an attempt to handle the possible correlated errors among similar datasets, Zwieback et al. (2012) suggested incorporating "expert knowledge" in the collocation procedure. Pan et al. (2015) proposed to adapt the previously described procedure to separate the structural and non-structural components of the errors under the assumption that similar datasets can be grouped, and that they are characterized by the same structural errors, whereas non-structural errors are uncorrelated within the same group. Grouping is defined *a-priori* based on the expert opinion on the similarity among products.

In this procedure, the collocation is performed on two steps (hence the name multi-step collocation, MC): first, the collocation is performed within each group, to derive an in-group optimal merge, and then a final merge is performed at the cross-group step. In our case, the similarity in modelling scheme is used as the *sole* factor to decide on the grouping, obtaining three groups: 1) a water balance (WB) group including the BIG BANG, LISFLOOD and LSA SAF datasets, 2) a residual surface energy balance (SEB) group including SSEBop and ALEXI datasets, are 3) a final separate group with the MOD16 product as the single member. The first step collocation applies a TC (Eq. (5) with $K = 3$) to the WB datasets to derive the errors associated to each model compared to the in-group truth (which differs from the real truth due to the similarities in the three models), and a two-component collocation of the two SEB products (Eq. (5) with $K = 2$), which corresponds to consider the simple

average as the optimal in-group truth (see Pan et al., 2015). At the second (cross-group) step, a TC is performed between the WB and SEB in-group optimal merges, and MOD16. The total errors are then computed as a sum of the errors in the two steps (structural and non-structural errors). This error assessment procedure corresponds to the graphical depiction in the Hilbert space reported in Fig. 3 of Pan et al. (2015) (with Group (1) = WB, Group (2) = SEB and Group (3) = MOD16), and it is summarized for our application by the scheme reported in Fig. 2. The error estimates obtained with this procedure can be used in Eq. (6) to derive a set of weighting factors to be used in Eq. (3).

The same weighting factors can be used to compute the weighted mean absolute deviation (wMAD) of the best-guess anomaly in Eq. (3), which corresponds to our estimate of the spread, ε^{z}_{my}, in Eq. (2). It is worth highlighting that the weighting factors computed with Eq. (6) are based on the average model errors (see Eq. (5)), and they are valid only when all six datasets are simultaneously available. Several factors may contribute to the absence of one of more datasets for a given month. As an example, the MOD16 product is systematically masked over large urban areas and inland water, ALEXI and LSA SAF are sporadically affected by the lack of data due to cloud coverage, and only LISFLOOD and BIG BANG are continuously available throughout the nineties. For the months in which some of the datasets are not available, the weighting factors of the remaining datasets are proportionally rescaled to add up to the unity.

3.2. Estimation of the climatology

The role of μ_m is to rescale the anomalies, modelled as described in section 3.1, to actual ET values (see Eq. (1)). This rescaling should account for the fact that ET datasets retrieved with different methodologies and modelling hypotheses may be characterized by rather different magnitudes due to systematic biases, and they may also cover different time periods within the target time window (1991–2020 in this study).

In absence of a well-established reference dataset, the terms characterizing the climatology in Eq. (1) are here estimated following an ensemble weighted median approach. For a given month, m, all the monthly data available from all the datasets are pooled together in a single ensemble of size N_m:

$$N_m = \sum_{k=1}^{K} n_m^k \tag{7}$$

where n_m^k is the number of valid data available for the k-th dataset, and K is the number of available datasets (equal to 6 in this study case). It is worth noting that for those datasets with missing data, n_m^k may vary from one month to another; otherwise, it will coincide with the number of available years.

The value of μ_m for each month is computed as a weighted median of the N_m ensemble members, with the weight of each member depending on two factors: 1) the weight of the specific dataset (from which the member is extracted), and 2) the number of available years (for the given month) in the specific dataset. In absence of *a-priori* information on the overall absolute accuracy of each dataset, a uniform weight can be given for all the datasets (i.e., $W_k \equiv W = 1/K$) (factor 1). Successively, the weight of each ensemble member is computed by dividing the overall weight of the corresponding dataset (W_k) to the number of available years for that specific dataset ($w^k_{my} = W_k/n^k_m$) (factor 2). This approach allows preserving the overall weight of each dataset, by avoiding to overweight the median toward datasets with more available years.

The spread of the climatological estimates, ε^{μ}_m, to be used in Eq. (2), can be computed as a weighted mean absolute deviation, in analogy to what is described in section 3.1 for the anomalies. The weights used for the spread are the same used for the computation of μ_m, as described in the previous paragraph (w^k_{my}). In this application, the climatological values obtained using a "uniform" weight for each dataset ($W = 1/6$) will be referred to as μ_u and ε^{μ}_u (the subscript m identifying the month is neglected hereafter for the sake of clarity). As an alternative method, the same weighting factor can be given to the three groups defined in section 3.1 ($W_{WB} = W_{SEB} = W_{MOD16} = 1/3$) rather than to the six datasets, hence obtaining a "variable" weight for each dataset. The statistics obtained with this approach, aiming at reducing the overfitting over similar datasets, will be referred to as μ_v and ε^{μ}_v. In both the uniform and variable cases ("u" and "v"), a single climatology is obtained, which will be used for the entire period independently from the datasets available in a specific year. This solution aims at improving the internal temporal consistency of the final ensemble ET series.

4. Results

The merging strategy described in section 3 treats the anomalies and the climatology separately; hence, the results related to these two components are discussed in two different sub-sections. Successively,

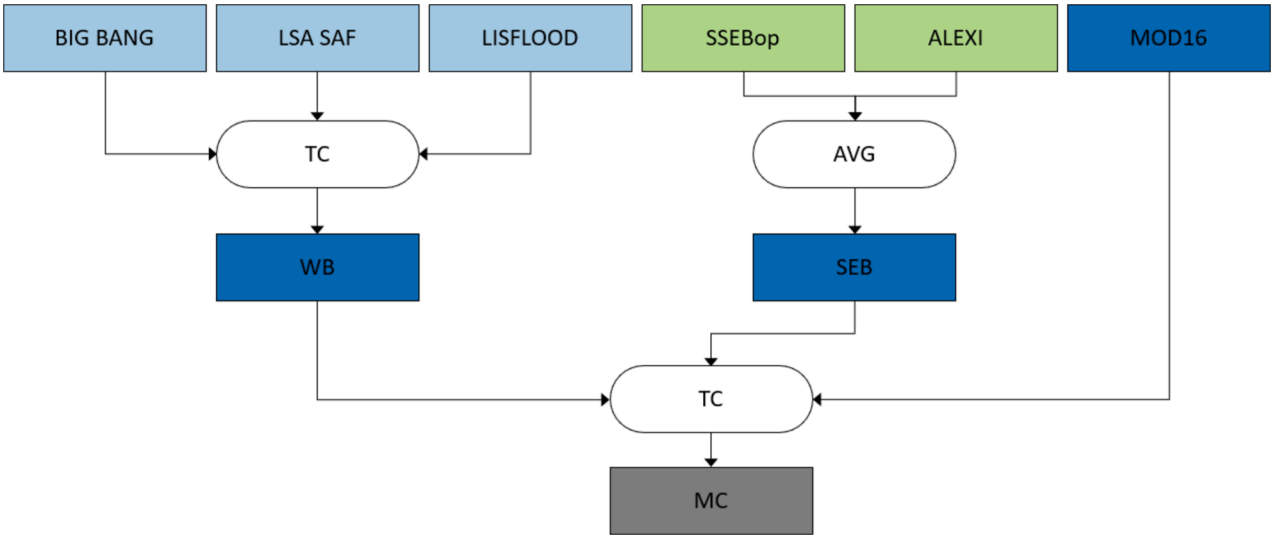


Fig. 2. Scheme of the multi-step collocation (MC) procedure. At the first step, structural errors are evaluated by combining the three water balance (WB) products (group (1) with a triple collocation (TC), and the two surface energy balance (SEB) datasets (group (2) with a simple average. At the second step, the in-group truths (WB, SEB and MOD16) are combined with a TC procedure to assess the non-structural errors.

the behavior of the merged product is analyzed, and compared against the EC observations.

4.1. Application of the multi-step collocation method

The multi-step collocation is applied to the ET anomalies, computed on the common overlapping period 2011–2020. Since one of the key hypotheses of any collocation is the linear relationship between the datasets and the (unknown) true status, a preliminary correlation analysis is performed to analyse the relationship between the six datasets. This analysis can also be used as a hint on their similarity, a valuable piece of information in MC analyses even if it is not directly representative of the hypothesis on zero cross-covariance in the errors at the base of MC.

The plot in Fig. 3 summarizes the pair-wise correlation among all six anomaly datasets by depicting the domain-average Pearson correlation coefficients (r). These results show an overall positive correlation for all the pairs (ranging on average between 0.11 and 0.37), with the highest values observed for the two pairs BIG BANG-LISFLOOD (0.37) and BIG BANG-LSA SAF (0.31). Correlation values are statistically significant ($p < 0.05$) in more than 50 % of the domain for all the pairs, and up to almost 90 % of the domain for the pair BIG BANG-LISFLOOD (see values within the circles in Fig. 3). The three WB models have consistently high correlation between them, suggesting a strong resemblance in their outputs due to the analogies in modelling framework and key inputs (i.e., dependence in the errors). BIG BANG and LISFLOOD, in particular, are both forced with station-based meteorological data, and they are the two datasets with the largest similarity.

The grouping of WB models based on expert knowledge seems to be supported by these results; however, this is not the case for the SEB group. The SSEBop-ALEXI pair reporting about-average correlation (0.20) despite their similarity in the key factor capturing water stress (MODIS LST) and in the modelling approach (residual SEB). This result may be mostly explained by the low correlation of SSEBop with all the datasets, while both SEB models have low correlation with LISFLOOD. The MOD16 product is well correlated with all the other products, with

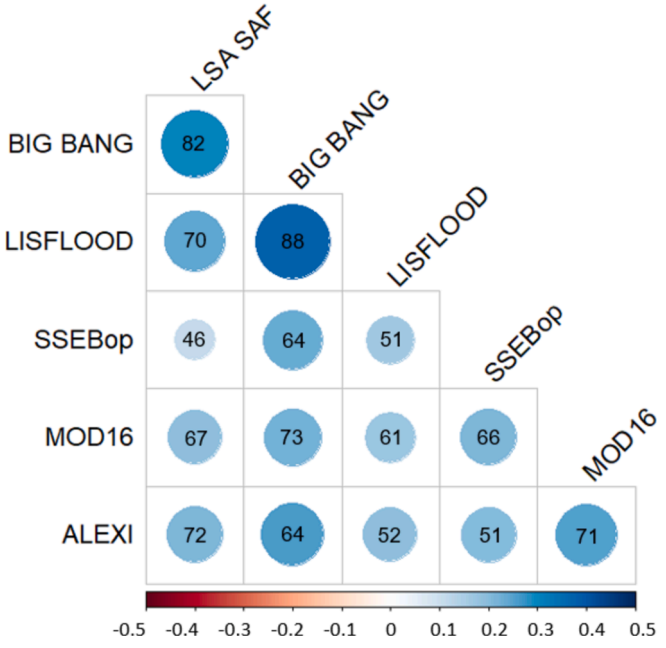


Fig. 3. Summary of the pair-wise correlation analysis performed between different ET anomaly datasets. The circles represent the domain-average Pearson correlation coefficient, with both size and color proportional to the magnitude. The number inside each circle represents the fraction of the domain with statistically significant ($p < 0.05$) positive correlation values.

no particular differences in the relationship with WB or SEB groups.

The application of the two-step MC procedure described in section 3.1 leads to the estimates of the weighting factors (ω^k) to be used in Eq. (2). As these weights are derived from average errors (Eq. (5)), they represent the local contribute of each dataset to the anomalies in the merged ensemble product. The spatial distribution of these weighting factors is depicted in Fig. 4 for the three groups WB, SEB and MOD16. In the upper-row panels (Fig. 4a-c), the color scale is chosen to represent about-average weighting factors ($\omega = 1/3 = 0.3333$) in yellow, and higher- or lower-than-average values in green and red shades, respectively. The lower-row maps depict instead the relative weight of each dataset to the total WB weights.

Some clear patterns can be observed in the weighting factor maps, with higher-than-average ω values over many areas in the north for WB (Fig. 4a), and over the south and Sicily for MOD16 (Fig. 4c). In the north-central regions, SEB has slightly above average weights, especially over Emilia-Romagna and western Piedmont (Fig. 4b). The split of the WB weights among the three datasets highlights the overall larger contribution of LISFLOOD across the domain, even if this dataset seems to have the lowest weight among the three models in the areas where WB contributes the most (e.g., Alps and Apennines). Over these regions, the LSA SAF seems to have high relative weight (Fig. 4e). The overall low relative weight of BIG BANG across the domain can be explained by the similarities with LISFLOOD. In this case, BIG BANG is not adding further information to the ensemble. Noteworthy is the case of the two SEB models, which are characterized by the same weighting factors due to the simple average procedure applied at the first step. This procedure seems to level out the contribution of the two models (which are characterized by a relatively low correlation, see Fig. 3), resulting in about-average weighting factors in many areas of the domain (see yellow areas in Fig. 4b).

These results can be further analyzed in terms of deviations from a theoretical simple average approach (i.e., same weighting factor for all the groups, equal to 1/3). The difference between the actual weighting factors and the simple average is summarized by the histograms reported in Fig. 5. These data highlight how both WB and MOD16 tend to have a significant part of the domain with higher weights in the MC approach compared to the simple average (positive differences), which is especially true for MOD16 (large right skewness, Fig. 5c). In the case of the WB model these correspond to northern regions (Fig. 4a), while they are mostly in the south for MOD16 (Fig. 4c). The differences observed for the SEB weighting factors seem to be more symmetrical and closer to zero (i.e., simple average) compared to WB and MOD16, even if a slight tendency toward negative values (less weight than average) is observed, with most of the differences between 0 and −0.2. As previously mentioned, this result can be due to the simple average performed at the first-step between SSEBop and ALEXI within the SEB group. Instead, for the MOD16 the assumption of no in-group structural errors may partially explain the overall larger weights compared to the other two groups (underestimation of structural errors). It is worth highlighting how all the weights need to add up to the unity; hence lower than average weights of a given dataset may just be associated to particularly low error estimated by the MC for another dataset.

Following the structure of the multi-step collocation, an additional set of correlation analyses is performed between the anomalies obtained for the three groups after the first step. The results of these analyses are summarized in Table 2, showing similar values for all three pairs, which is particularly evident in the fraction of the domain with statistically significant positive correlation (almost 80 % in all three cases).

4.2. Computation of ET climatological references

The results reported in section 4.1 focus on the anomalies from the climatology; hence, any systematic differences in the magnitude of the datasets were preliminarily removed. Here the seasonal dynamics of the climatology are analyzed, and in particular those of the two sets of

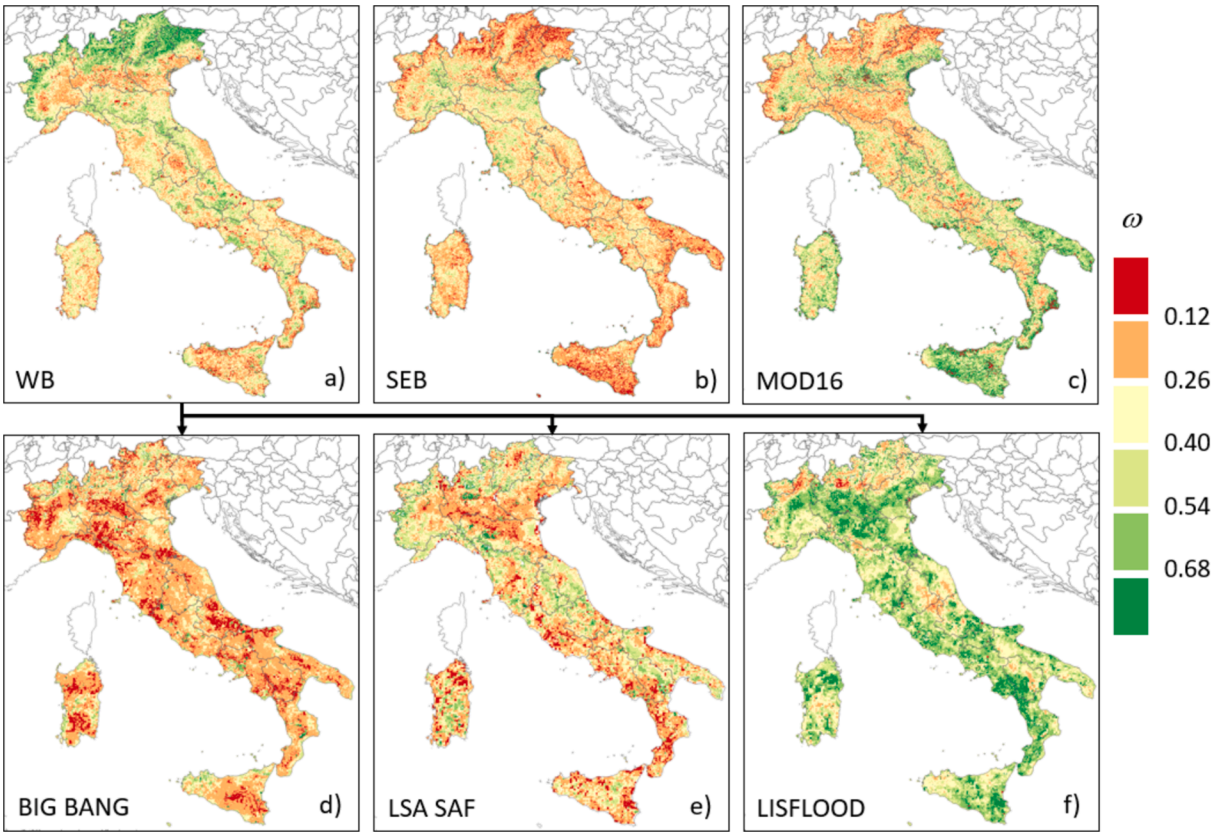


Fig. 4. Spatial distribution of the weighting factors derived for the three groups (panels a-c, see section 3.1) using the multi-step collocation (MC) procedure. The lower row panels (d-f) report the relative contribution of each dataset to the WB weights.

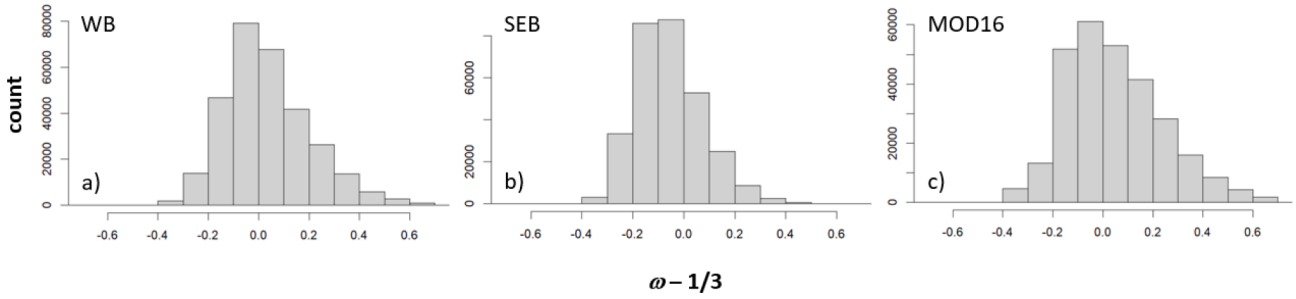


Fig. 5. Frequency distribution of the difference between the weighting factors obtained with the MC and the simple average (ω = 1/3). Positive values correspond to higher weights in the MC compared to the simple average.

Table 2
Summary statistics of the correlation analysis between the three model groups. The upper matrix reports the average (± standard deviation) Pearson coefficient, whereas the lower matrix reports the fraction of the domain with statistically significant ($p < 0.05$) positive values.

	MOD16	WB	SEB
MOD16	–	0.28 ± 0.39	0.32 ± 0.40
WB	76	–	0.38 ± 0.21
SEB	78	79	–

climatological references obtained with the uniform (subscript u) or variable (subscript v) weighting factors (as described in section 3.2). The spatiotemporal patterns of the climatological quantities are represented in the maps in Fig. 6, reporting the seasonal values (DJF = December-February; MAM = March-May; JJA = June-August; SON = September-November) for μ_v (expressed as mm/d). The general behaviors observed for these maps are also valid for the analogous μ_u maps (not shown).

Overall, the climatology of the median follows the expected seasonality of the area, with smaller ET during winter months (DJF) and higher values during summer (JJA). In winter, ET is mostly controlled by the radiative forcing, hence larger values are modelled in the south (between 0.5 and 1 mm/d) compared to the north (< 0.5 mm/d). On the contrary, water limited conditions arise during the warm summer, especially in the southern areas, and an inversion in the ET magnitude (larger values in the north compared to the south) can be observed. During summer (JJA), the median ET reaches values well-above 3 mm/d over most of the mountain regions, whereas it rarely exceeds 1.5 mm/d over most of the agricultural areas (see Fig. 1) in Sicily, Sardinia, and Apulia.

Even if the general patterns of the two climatological datasets are quite similar, some interesting results can be observed when the differences between the monthly maps are investigated, as summarized by

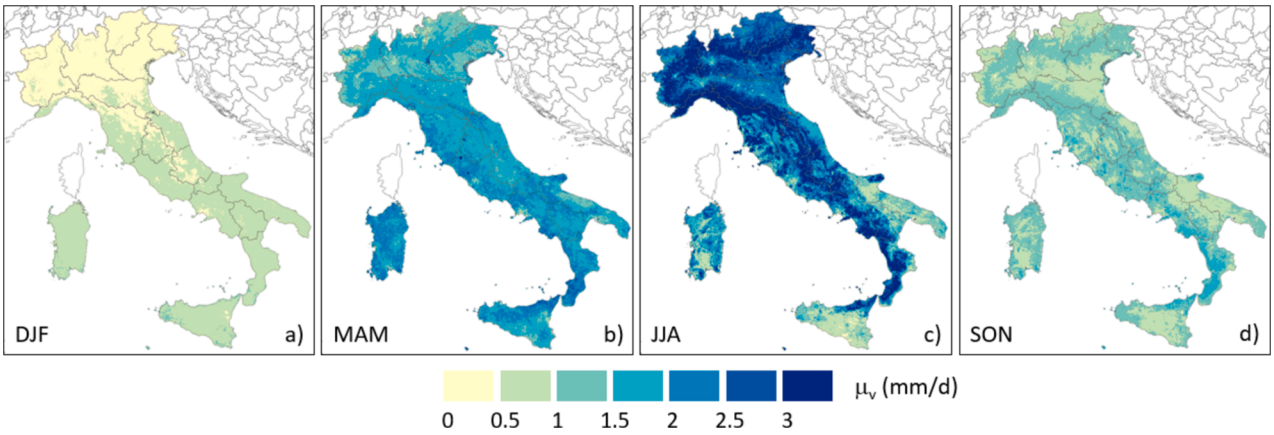


Fig. 6. Spatial distribution of the seasonal median values derived as a weighted ensemble of the six datasets. The variable weight is used in this case (μ_v, see section 3.2). The four seasons are defined as follow: DJF = December-February; MAM = March-May; JJA = June-August; SON = September-November.

the box plot reported in Fig. 6. Here, the differences computed on the monthly maps highlight overall higher μ_v values compared to μ_u (negative differences in Fig. 7), also with some clear seasonal variability. Larger ($\mu_u - \mu_v$) differences, on average, are observed for the cold months (November to March), albeit with a very limited spatial variability and magnitude. A larger spread in the differences is observed during the warm months (June to August), where the frequency of positive differences is not negligible anymore. These positive values are mostly located over agricultural land, whereas negative values in summer months are observed over forest areas. Generally, most of the differences between the two climatologies are driven by the tendency of WB models to return lower values compared to SEB and MOD16 (except for cropland during summer); hence, the "variable" method returns higher ET due to the lower weight given to the WB models. Overall, however, the observed differences are of limited magnitude, as the two climatologies are generally comparable in terms of spatio-temporal features.

4.3. Comparison with EC observations

The multi-model anomalies obtained with the weighting factors described in section 4.1 and the climatology derived in section 4.2 are combined to obtain the merged product over the entire study period and area. The monthly EC measurements described in section 2.1 are used to quantify the performance of this merged product, and the results obtained for the six base datasets are used as a further benchmark. In addition, the simple average of the six ET products (AVG) is also tested as a simpler merging strategy. The plots in Fig. 8 combine some

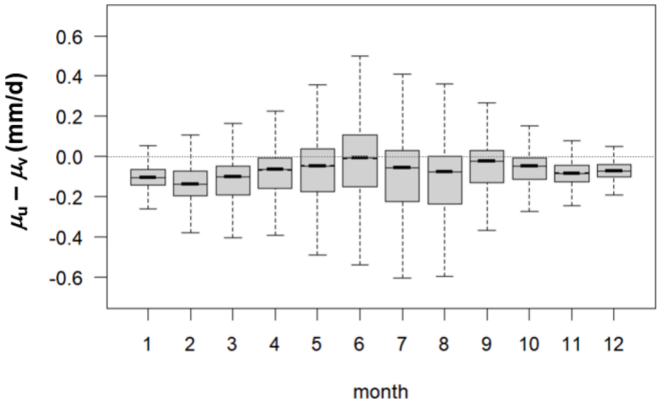


Fig. 7. Box plot depicting the difference between the two climatological datasets (subscript u and v, see section 3.2) in terms of monthly weighted median.

statistical metrics for all the sites quantifying the agreement (mean absolute difference, MAD, and relative difference, RD), systematic differences (BIAS), and correlation (Pearson correlation coefficient, r). These results show roughly similar agreement metrics for the six methods at monthly scale, with MAD ranging on average between 0.52 and 0.68 mm/d, and RD between 30 and 40 %. In term of BIAS, all the three WB models seems to systematically underestimate the EC fluxes (BIAS of about −0.35 mm/d), whereas ALEXI tends to overestimate by a similar quantity (0.2 mm/d). Despite the slightly coarser spatial resolution of some datasets (LSA SAF and ALEXI) compared to the target resolution of 1-km, the performance of these products does not seem to be negatively affected, given the overall similarities in all the statistical metrics.

Since all the WB models are characterized by negative biases, and given the observed predominance of negative differences in Fig. 7, we decided to adopt the "variable" climatology (μ_v, see section 3.2) for the comparison with the EC observations reported in Fig. 8 and for all the analyses hereafter. However, we have observed that this choice has only a minimal effect on the ensemble dataset behavior (not shown). Overall, the MC method displays good performance, with low differences (MAD = 0.47 ± 0.17 mm/d, RD = 27.9 ± 7.5 %), limited bias (BIAS = −0.17 ± 0.26 mm/d,) and high correlation (r = 0.83 ± 0.10). Given the similarity of all the six datasets in terms of MAD and RD (Fig. 8a and 8c), these metrics for both the AVG and MC datasets are only slightly lower than the ones for the dataset with the lowest differences (MOD16). A more marked difference can be observed on the correlation, where the performances of BIG BANG and SSEBop do not seem to negatively affect the result of the MC approach. In term of bias, while MC has lower BIAS (in absolute value) than most of the models, it does not outperform MOD16 on this limited set of sites.

It is worth mentioning that, albeit slightly, MC always outperform the simple average (AVG), and how the small negative bias observed in most of the models (except for ALEXI) is in line with the use of the Bowen ratio closure on the observation, which increases the ET compared to the actual (unclosed) observations. Another key result highlighted by Fig. 8 is the reduced inter-site variability (error bars in Fig. 8) of the MC compared to the single datasets, suggesting more uniform performance across the domain. Due to the small number of EC stations, and their limited spatial representativeness, it is not possible to extrapolate additional insight beyond a general successful implementation of the proposed ensemble methodology.

Indeed, if the MAD for each site is analyzed independently, as depicted in Fig. 9, no clear differences between the spatial patterns in the six based datasets and the merged product can be observed. Overall, more spatially homogeneous results are observed in MC (as already highlighted by the error bar in Fig. 8a), even if slightly larger errors are observed over PT1 compared to the other sites. Over this site, only

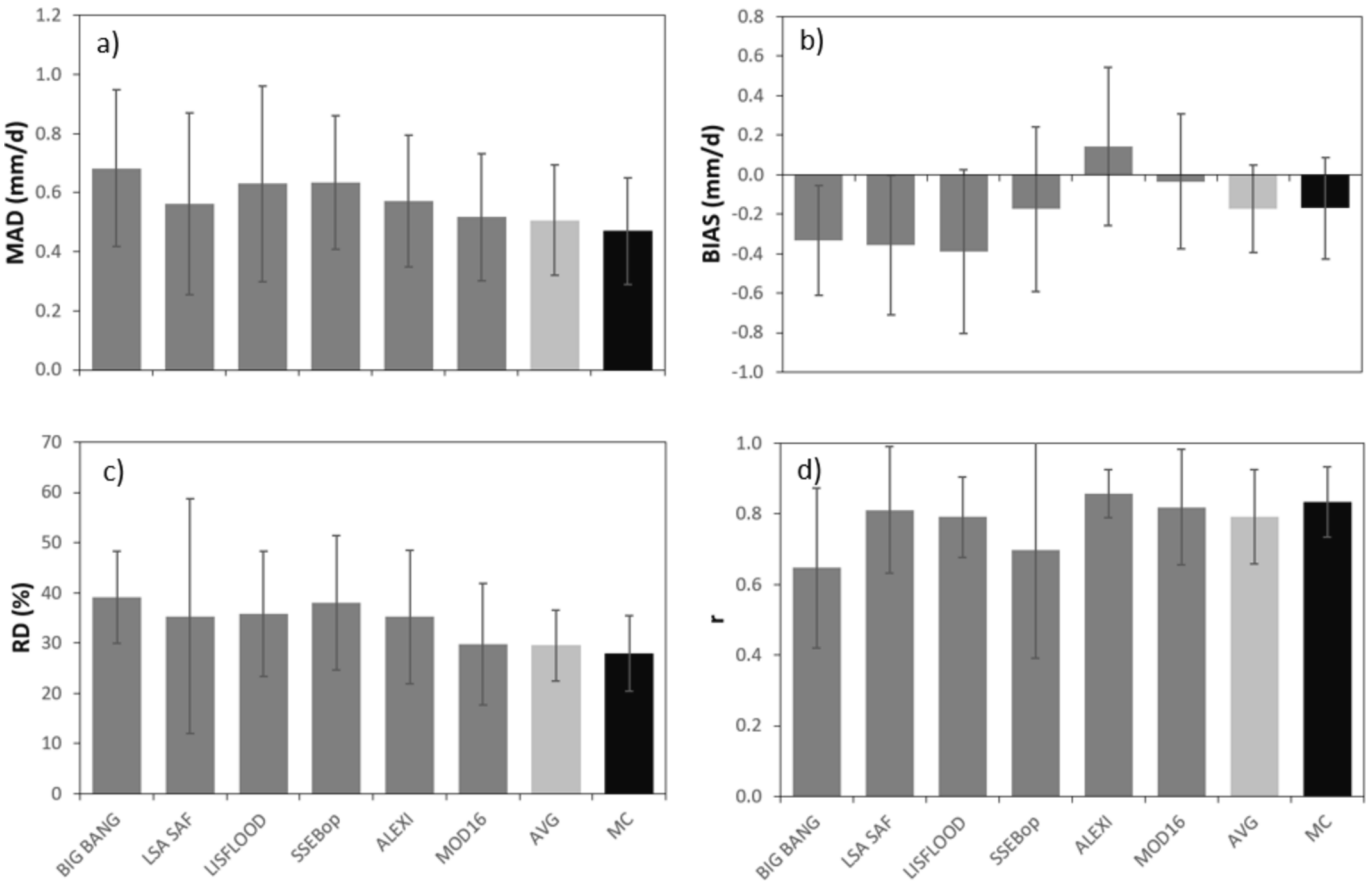


Fig. 8. Result of the intercomparison of the selected MC model with the EC measurements. Statistical metrics are contrasted with the ones obtained for each single dataset (see Table 1), as well as the simple average of the six datasets (AVG). Panel a) reports the mean absolute difference (MAD), panel b) the bias (BIAS), panel c) the relative difference (RD), and panel d) the Pearson correlation coefficient (r). Error bars show the intra-site variability of the metrics.

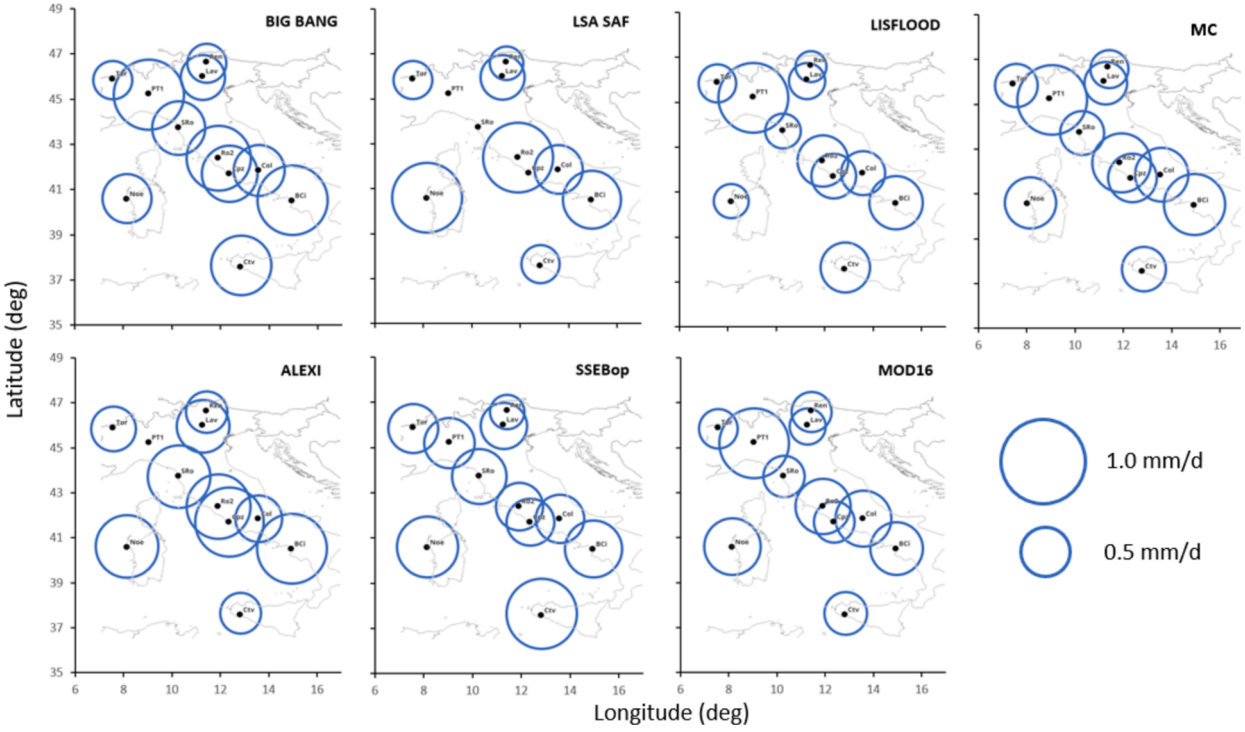


Fig. 9. Spatial distribution of the mean absolute difference (MAD) computed between the eddy covariance observations and the six based datasets and merged product (MC).

SSEBop seems to have a MAD around 0.5 mm/d, while higher MAD values are observed for BIG BANG, LISFLOOD and MOD16 (close to 1 mm/d). It is important to highlight how MAD values cannot be computed for some sites and datasets due to the absence of spatiotemporal overlap with the eddy covariance observations (e.g., ALEXI and LSA SAF over PT1).

4.4. Spread and temporal consistency of the ensemble estimates

Two of the main objectives of the study are to provide information on the spread and to ensure internal temporal consistency in the merged products. The spatio-temporal dynamic of the median climatology obtained with the "variable" method and its spread are summarized for the four macro-regions (see section 2.1) in the plots in Fig. 10. These plots highlight some key differences in the climatology of the four regions, with a peak ET that occurs later in the year moving northward, from May in the Insular (ITG; Sardinia and Sicily) region to July in the North (ITC + ITH) region. The temporal dynamic follows the irradiation curve only for the North region, whereas the effect of water stress is partially visible between June and September in both the Central and the South regions, and it is even more marked over the Insular region.

Overall, a larger spread is observed in the summer months compared to winter months for all four regions. On the one hand, this is somewhat explained by the larger magnitude of the ET fluxes during the months with high irradiation, which is true for the North, and partially the Central, regions. During this period, differences in meteorological forcing may be the major source of the discrepancy among the models. On the other hand, the very marked spread in South and Insular regions during the period July-September is most likely related to the differences in water stress modelling, with a spread that can become comparable to the best-guess itself (e.g., for the Insular region in August: $\mu = 1.15$ mm/d and $\varepsilon^{\mu} = 0.55$ mm/d).

Since the climatology incorporates the systematic biases among the six datasets, it is observed that ε^{μ} represents a significant fraction of the total spread, ε (see Eq. (2)). The data in Table 3 report the results of the linear regressions between ε and ε^{μ} values averaged over the four seasons. The regression lines (forced though the origin) have values ranging between 0.8 and 0.86, suggesting that at least 80 % of the total spread is roughly incorporated in the climatological spread.

Table 3
Summary of the relationship between total, ε_{my}, and climatological, ε^{μ}_{m}, spreads averaged over the four seasons: December-February (DJF), March-May (MAM), June-August (JJA), and September-November (SON).

Season	Intercept	Slope	Pearson	Slope (Intercept = 0)
DJF	0.04	0.66	0.97	0.80
MAM	0.09	0.68	0.98	0.84
JJA	0.13	0.70	0.98	0.83
SON	0.08	0.69	0.96	0.86

For the four seasons, the maps in Fig. 11 report the spatial distribution of the season-average total spread, as both actual value (upper row, panels a-d) and coefficient of variation, CV = ε/μ (lower row, panels e-h). In terms of actual value, the results follow the behavior reported in Fig. 10 for the climatological component, with larger spread during summer months (JJA, Fig. 11c) and lower spread in the winter (DJF, Fig. 11a). The largest values ($\varepsilon > 1$ mm/d) are observed in the South and Insular regions, especially over mountain forested areas (e.g., Nebrodi in Sicily and Apennines in Calabria).

In relative terms, the CV values are higher during winter in mountain regions (Fig. 11e), likely due to the small magnitude of the climatological median, as well as during summer over Sicily, Sardinia and Apulia (Fig. 11g). The high value of CV in summer are not limited to forested areas, as for the absolute ε, but they are distributed across the entire southern Italy.

One final set of plots aims at showing the internal temporal consistency of the time series obtained with the adopted merging strategy. The plots in Fig. 12 show the time series of total annual ET obtained with the MC method and the simple average (AVG) of the six datasets. These plots show how the different temporal coverage of the six datasets causes some artificial discontinuities in the simple average dataset (grey lines in Fig. 12), which are not visible in the MC product (black lines in Fig. 12). Those discontinuities are mostly located around the year of introduction of new datasets (vertical dotted lines in Fig. 12). The similarity between the two datasets is much more marked in the last part of the period (after

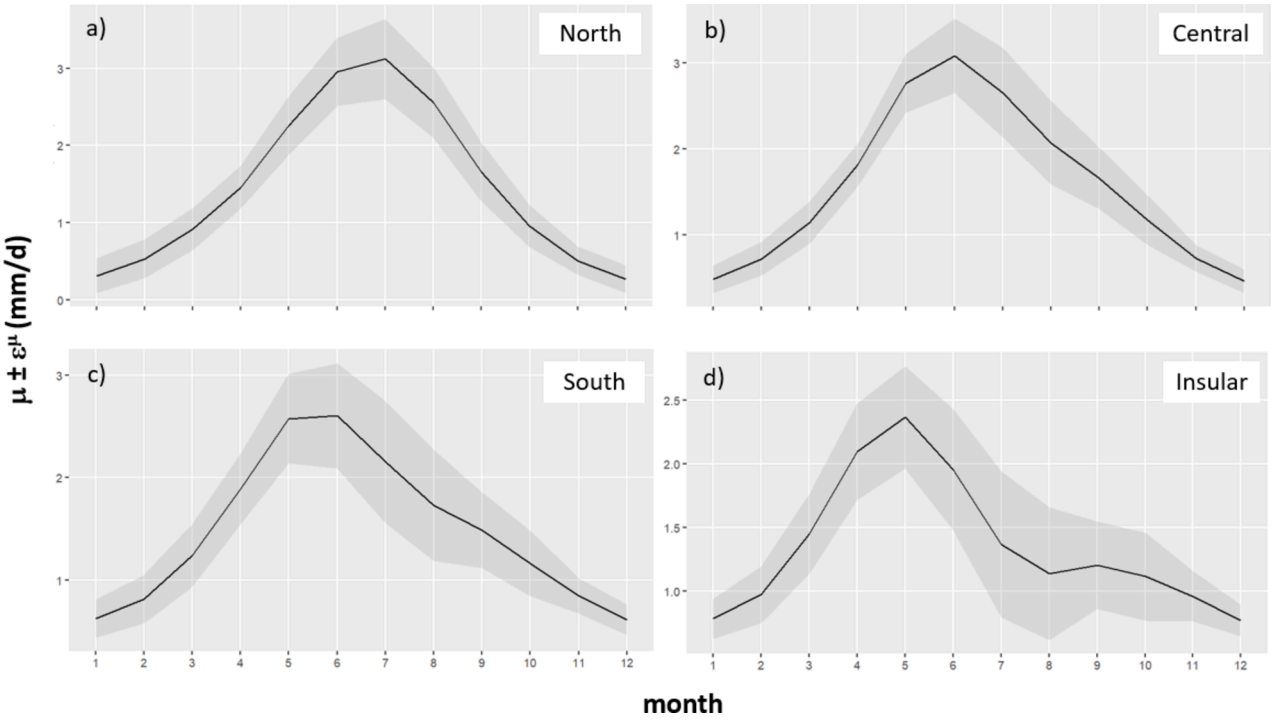


Fig. 10. Annual cycle of the climatology (black line) and its spread (grey shade) for the four macro regions: a) North, b) Central, c) South, and d) Insular.

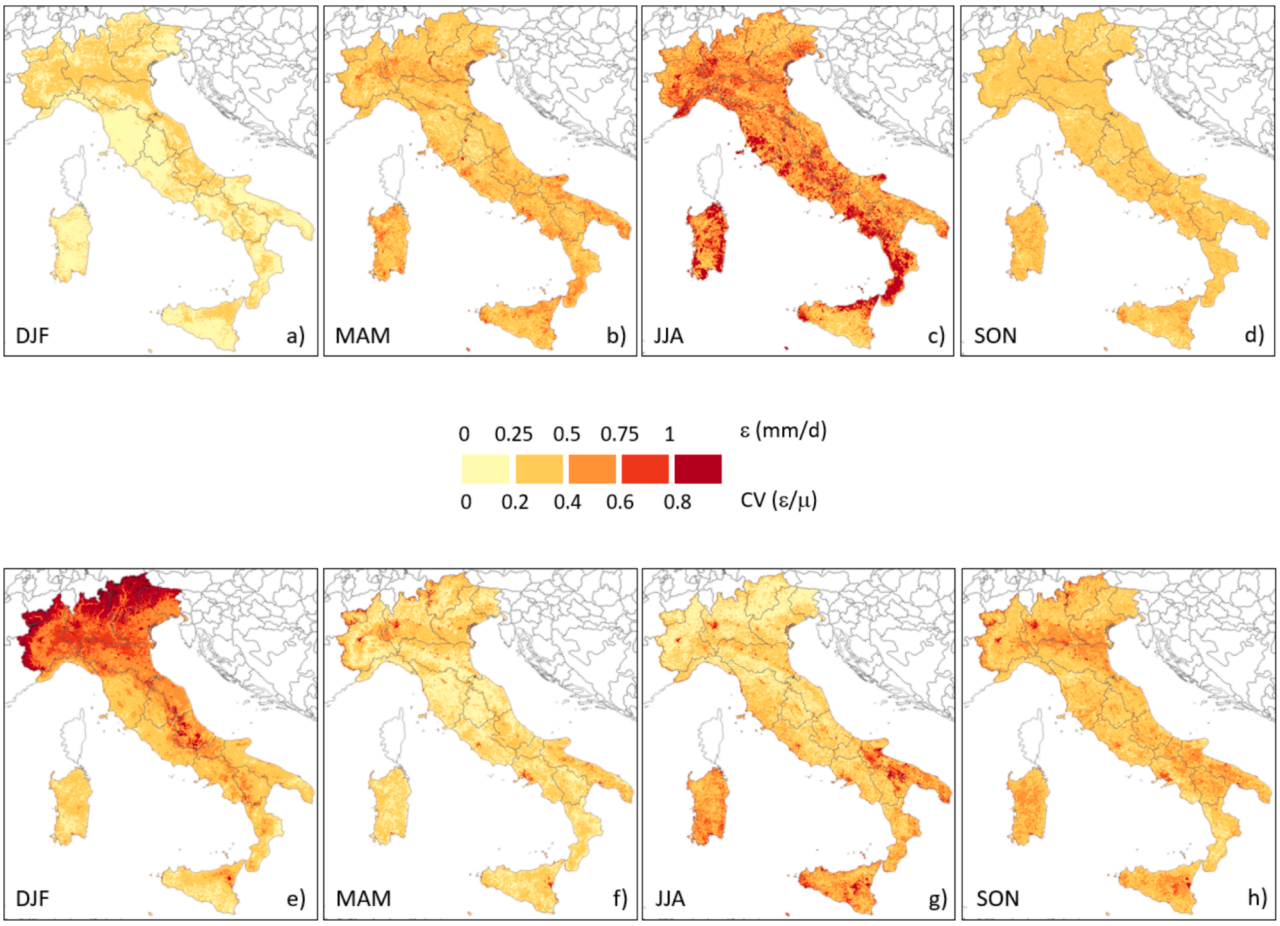


Fig. 11. Spatial distribution of the season-average total spread as actual value (ε, panels a-d) and coefficient of variation (CV, panels e-h).

2011) when all six datasets are available. However, systematic differences can be still observed due to the different weighting factors use by the two approaches (e.g., higher values for MC compared to AVG in panel d and lower values in panel c).

5. Discussion

The principle behind the ensemble modelling is the notion that estimates from an ensemble of models are usually equally or more accurate than the individual assessments, as supported by many applications in climate science and hydrology (e.g., Arsenault et al., 2015; Kirtman et al., 2014; Melton et al., 2022; Thompson, 1977). In this framework, similarities among datasets play a major role, since on the one hand the ensemble estimates tend to gravitate toward converging evidence, but on the other hand estimations may be biased by modelling approaches adopting similar schemes or forcing.

In triple collocation (TC) applications, the issue of cross-correlation between product errors is intrinsically solved by applying the approach only on three datasets based on different conceptualizations (e.g., land surface models, active and passive microwave for soil moisture, see Gruber et al., 2016). When the number of datasets increases, the assumption that all the pairs have zero cross-correlation becomes increasingly unrealistic. The adopted multi-step expert-based approach solves this issue similarly to classical TC, by grouping the data based on their similarity in modelling scheme. Following this approach, the number of groups is mostly dictated by the available modeling frameworks in the scientific literature. In this regard, direct ET estimations from soil moisture microwave data may represent an additional group, but no significant expansion in the number of possible groups is foreseen. Even if the number of groups may not be significantly increased in further applications based on different datasets, the association of a dataset to a group or another is still subject to some arbitrariness and potential refinements.

In this study, we assumed that similarity is mostly driven by the modelling scheme, with WB approaches simulating the changes in potential ET from soil moisture-based water stress factors, SEB approach simulating the stress through the effects on LST, whereas the MOD16 approach focuses on water vapor deficit. This schematization is based on the assumption that the modelling of ET in absence of water stress is simpler, as confirmed by the reduced spread over northern regions where stress is limited or even absent. The large spread over southern and insular Italy suggests instead major differences cause by different water stress scheme. However, differences between models belonging to the same group are also documented in the literature, mainly driven by differences in the main meteorological forcings (Vinukollu et al., 2011). These differences may cause discrepancies that can be comparable to the inter-group differences (Zhang et al., 2020).

Even if a proper estimation of the cross-covariance between the errors in each dataset is not possible, due to the lack of information on the true status (Gruber et al., 2016), the use of an expert-based approach represents a step forward compared to neglecting the existence of cross-covariance and directly performing a multiple collocation on all six datasets. However, insight obtained from pair-wise correlation analyses performed on anomalies should be used to derive information on the agreement between models, since systematic differences and obvious relationships (e.g., due to seasonality in the observed quantities) are removed beforehand. In this study, Person correlation coefficients between ET anomalies are generally positive, but overall lower than values observed in similar studies on soil moisture (Cammalleri et al., 2015; Cammalleri et al., 2017). They are, however, still statistically significant

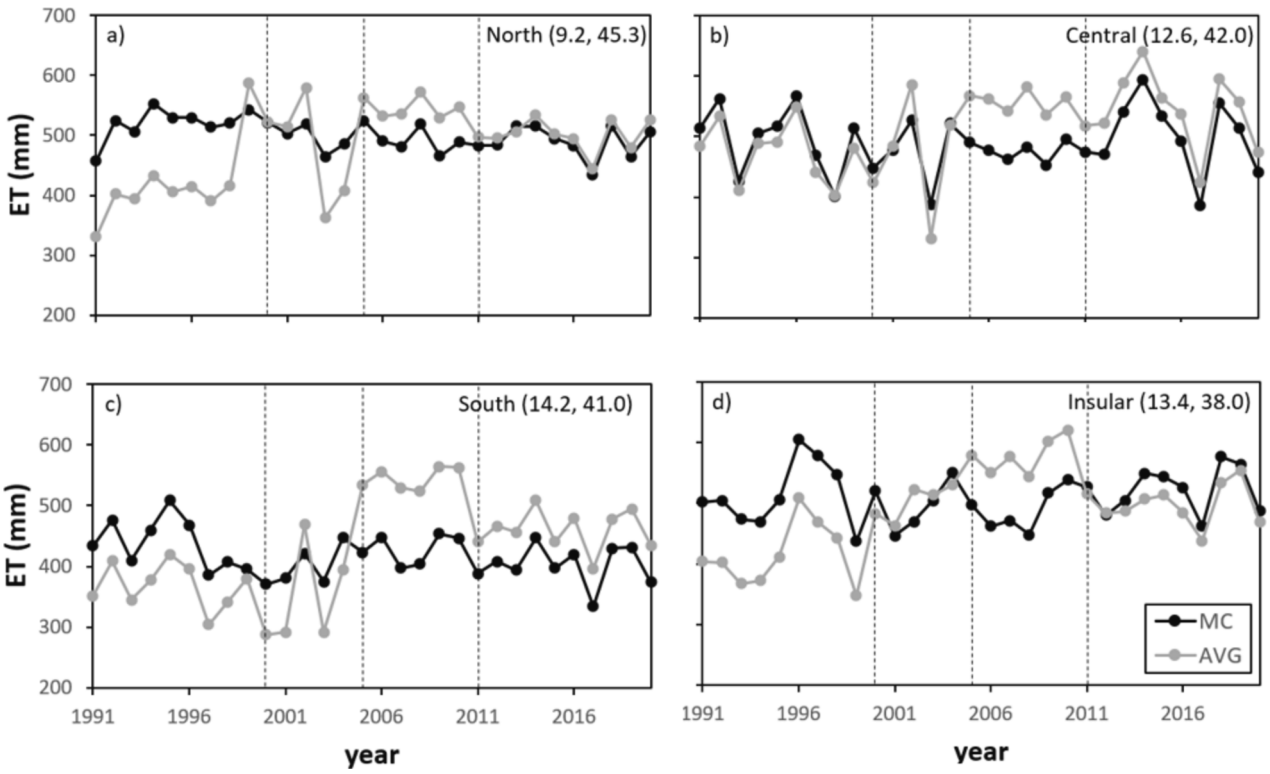


Fig. 12. Time series of total annual ET according to the MC method and the simple average (AVG) for four example sites in different macro regions (see Fig. 10). The vertical dotted lines represent the year of introduction of new datasets.

in most of the cases. This result may be explained by the fact that ET is influenced by variations in water availability (i.e., soil moisture) but also in the atmospheric water demand controlled by other meteorological (i.e., radiation and temperature) and environmental (i.e., vegetation type and amount) factors.

Differences among the various pairs show a tendency of the water balance (WB) datasets to agree more strongly. Since the main forcing of WB approaches is precipitation, the overall consistency between reanalysis and gridded rainfall datasets over Europe and Italy (i.e., Adinolfi et al., 2023; Hassler and Lauer, 2021) may explain the general agreement in modelled ET. This overall agreement supports the suggestion of Pan et al. (2015) to account for this similarity in the ensemble strategy by separately quantifying the structural and non-structural errors.

In contrast, the agreement between the anomalies modelled by the two thermal-based surface energy balance (SEB) models is not particularly high compared to the other pairs, despite the use of a similar MODIS-based LST dataset. This result suggests that similarity in LST data may be hidden by the strong difference in some assumptions adopted by the two modelling schemes, as well as different meteorological forcings. Also, while the two models were extensively validated over the US (Volk et al., 2024), performances over Europe are not equally known. Indeed, large differences in estimates from different SEB models have been observed in some intercomparison studies, albeit at higher spatial resolution and over shorter temporal scales (e.g., Acharya and Sharma, 2021; Bhattarai et al., 2016; Hu et al., 2023; Sun et al., 2019). Recently, De Santis et al. (2022) analyzed different remote-sensing based ET products over Italy, highlighting various degree of agreement with EC observations. These considerations suggest that the inclusion of expert knowledge in the ensemble modelling is a sound practical approach, following the common decision-making practice to account for the level of information of each player in taking decisions (Branzei et al., 2001). However, simple theoretical similarity among modelling schemes may not be enough to capture expected structural errors, and that additional factors (i.e., consistency in meteorological forcing or other inputs) should be also taken into account for a more refined definition of inter-model consistency.

A critical analysis of the weighting factor maps (Fig. 4) obtained with the MC approach highlights some clear spatial patterns, with large regions showing similar weights for a specific model. In particular, it is possible to observe how the approaches based on gridded meteorological data (WB methods, LISFLOOD and BIG BANG in particular) have high weights in the north, whereas methodologies indirectly assessing the effect of water stress have high weights in the south (especially MOD16). These patterns seem to resemble similar outcomes observed in Cammalleri et al. (2017) for soil moisture, where data rich areas are well modelled by water balance approaches, whereas remote sensing schemes are better suited for dry areas. On the one hand, good performances of the MOD16 product have been observed by Autovino et al. (2016) in Sicily, and by Ruggieri et al. (2021) in Campania. On the other hand, Castelli (2021) reported that both MOD16 and SSEBop may not be accurate enough over the Alps. Braca et al. (2021) highlighted an overall agreement between BIG BANG and MOD16 at annual scale in Italy, even if this result is due to a compensation effect on the opposite differences observed during winter and summer months. The consistency of the observed patterns in weighting factors with the behaviors observed in these past experiences further supports the successful implementation of the combining strategy.

Another key component of the merging strategy is the definition of a proper climatology. This represents a peculiar issue of merging ET products, since the actual magnitude of the fluxes is fundamental in defining the water use, but a reference dataset is missing. A clear outcome of this study is the main role that the climatology plays on the overall accuracy of the ensemble outputs, likely because intra-model systematic biases are resolved in this component of the ensemble modelling. This is also reflected in the contribution of climatology spread to the overall ensemble spread, suggesting that the inclusion of information on the spread should be a key feature of any ensemble strategy to highlight where and how improvements can be made. In our

application, the obtained optimal climatology was mostly dictated by the absolute bias of the six base products, and the systematic underestimation of most of the datasets (especially the WB ones). It is worth remarking how, on average, biases against in-situ data were limited for all the datasets (< 0.4 mm/d) and that systematic errors in ground observations can also be present due to energy unbalance and closure strategies (Allen et al., 2011). These two factors suggest that only limited improvements may be achieved by advancing the anomaly-combining method (i.e., collocation) without ensuring a reliable base climatology.

The comparison of the merged product against in-situ data confirms the good performances of the approach, with errors comparable with other similar studies (e.g., Li et al., 2023) and limited systematic biases. As the six base datasets are characterized by relatively similar accuracy, improvement over the single estimates a rather limited. However, these improvements are still significant as they are systematic and accompanied by a reduced inter-site variability (i.e., more homogeneous performances). The overall improvement of the ensemble method compared to the estimates of the single models seems to confirm the positive effects already observed in other studies. Over South America, Melo et al. (2021) highlighted how no model outperformed the others for all conditions in comparing four ET models, whereas Bai (2023) suggested that it is not recommended to use a single model to simulate ET over China, despite the good consistency observed on nationwide multi-year average ET by three remote sensing models. Recently, an ensemble strategy has been adopted over the western US for field-scale ET assessments as part of the OpenET initiative (Melton et al., 2022). While Beck et al. (2017) showed the value of a multi-model ensemble for runoff estimation, and how even including less-accurate models did not severely degrade the performance. This is particularly true for the MC-based strategies, where accuracy is directly accounted through the weighting factor.

Still, a proper quantification of the effects of the merging approach is not always straightforward, and strategy to achieve further improvements may not be as easy to identify. Improvement in the accuracy of a single model may not always result in a better ensemble outcome, as sometimes compensation effects may be in place (e.g., a model that overestimates may compensate the underestimation of another model). A clear example in this study is represented by the accuracy in the areas with large spread in ET estimates over southern Italy, where an improvement in ET accuracy may be obtained by simply adopting a more robust reference climatology, even without changes in inter-annual variability or ensemble scheme.

Beyond the observed good accuracy of the multi-model ensemble compared to the single datasets, meaningful considerations on the efficiency of the proposed combining strategies are limited by the scarce number of ground observations available in this study. In this regard, accuracy and correlation are higher compared to a simple average but not of a significant margin. Wang et al. (2015) already observed how more observation networks and better quality-controls can be crucial in improving ET estimates, and how mismatching in spatial scales and heterogeneity may substantially decrease correlation between observed and modelled data. Selection strategies that heavily rely on the outcomes on a limited number of ground observations may lead to misleading results, and for this reason qualitative analyses on the reliability of the spatial patterns may also be useful. Recently, Alfieri et al. (2022) applied GLEAM (Global Land Evaporation Amsterdam Model, Miralles et al., 2011) over the Po River basin, showing high correlation of modelled daily ET with data collected over the FluxNet "Tor" station. In our analysis, similar performances were obtained for all six analyzed datasets on this same EC station, suggesting a very strong agreement over this area that cannot be directly extrapolated for the entire national territory (as highlighted by the results observed on the other test sites). While the comparison with ground observations is encouraging on the reliability of the proposed merging strategy over the study area, the obtained results further stress on the importance of establishing a network of robust ET records to better strength the connection between ET estimates and the reality.

6. Summary and conclusions

The reconstruction of a climatological dataset of actual evapotranspiration (ET) for the Italian territory has proven to be challenging due to the absence of a clear common reference dataset, the lack of in-situ scale-appropriate observations, and discrepancies between different modelling schemes in terms of meteorological forcing and methods used to reproduce the effects of water stress. The proposed ensemble method confirms the possibility to obtain estimates that are, on average, better than the individual models, with good accuracy and low bias compared to a (limited) set of ground observations, and (most importantly) a more uniform performance across the domain.

In addition to the best-guess estimates, the ET dataset described here is accompanied by an assessment of the spread, which was proven to be useful for understanding the agreement among the base datasets. This variable highlights the important role of quantifying the spread of the climatology, and the need to properly consider this spread when searching for potential improvements on the methodology. In this study case, the spread seems to be more marked in areas characterized by high water deficit conditions (mostly modelled in southern Italy during summer), further stressing how one of the major sources of discrepancy in actual ET modelling is the quantification of water stress and the accuracy in the precipitation input.

The multi-step collocation method adopted for the combination of the anomalies has the advantage of being able to handle any number of input datasets, beyond the simple three of the triple collocation, expanding the possibility for merging to a high number of ET datasets. With the increasing availability of freely available operational datasets, the need to account for the cross-covariance between dataset errors will become a more urgent task. In this study, the issue has been tackled by including expert knowledge in designing the ensemble procedure, which returned reliable patterns in weighting factors. The relatively high correlation observed between water balance models seems to suggest how these models may belong to the same grouping sharing similarities; however, a similar agreement was not observed between the two residual surface energy balance models, suggesting that expert knowledge may need to go beyond simple similarities in modelling scheme, but also account for sensitivity to key inputs (e.g., LST) and similarity in other input variables.

Moving forward, the adoption of an ensemble approach has the advantage to remove any confusion for the potential users on which dataset should be adopted as a reference, as it produces an optimal synthesis of the various estimates. While the improvement of the combined datasets over the single methods is clear, and in agreement with many studies highlighting the value of ensemble modelling, the limited dataset of ground observations available in this study area do not allow to further corroborate on the added value of the expert knowledge-based ensemble scheme here tested. This limitation urges for a more systematical collection of ground observation ET data.

Author contributions

CaC designed the experiments, with inputs from MM and ChC. CaC developed the codes and performed the analyses, with inputs from MA, CH and YY for the ALEXI data and PS for the LISFLOOD data. CaC prepared the manuscript, which was revised by all co-authors.

CRediT authorship contribution statement

C. Cammalleri: Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **M.C. Anderson:** Writing – review & editing, Data curation. **C. Corbari:** Writing – review & editing, Methodology. **Y. Yang:** Writing – review & editing, Methodology, Data curation. **C.R. Hain:** Writing – review & editing, Data curation. **P.**

Salamon: Writing – review & editing, Data curation. **M. Mancini:** Writing – review & editing, Project administration, Funding acquisition.

Declaration of competing interest

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

Acknowledgments

This work was partially founded by "Programma di ricerca, CN_00000022 "National Research Centre for Agricultural Tecnologies (Agritech)" finanziato dal Decreto Direttoriale di concessione del finanziamento n. 1032 del 17.06.2022 a valere sulle risorse del PNRR MUR – M4C2 – Investimento 1.4 - Avviso "Centri Nazionali" - D.D. n. 3138 del 16/12/2021 - Spoke 3 Enabling Technologies and sustainable strategies for the smart management of agricultural system and their environmental impact".

We acknowledge the key role of the open access datasets made available by FLUXNET, ISPRA, EUMETSAT, USGS, and Copernicus CDS. In particular, we would like to thank Cinzia Mazetti at ECMWF and Stefania Grimaldi at the EC-JRC Floods group for the LISFLOOD data, and Giuseppe Ciraolo, Dario De Caro, and Matteo Ippolito at the University of Palermo for providing the additional eddy covariance data.

Data availability

The ensemble dataset produced within this research is made publicly available via Zenodo (doi: 10.5281/zenodo.8359213). Codes used for the processing can be shared upon request to the main author.

References

Acharya, B., Sharma, V., 2021. Comparison of satellite driven surface energy balance models in estimating crop evapotranspiration in semi-arid to arid inter-mountain region. Remote Sens. 13 (9), 1822. https://doi.org/10.3390/rs13091822.

Adinolfi, M., Raffa, M., Reder, A., Mercogliano, P., 2023. Clim. Dyn. https://doi.org/10.1007/s00382-023-06803-w.

Alfieri, L., Avanzi, F., Delogu, F., Gabellani, S., Bruno, G., Campo, L., Libertino, A., Massari, C., Tarpanelli, A., Rains, D., Miralles, D., Quast, R., Vreugdenhil, M., Wu, H., Brocca, L., 2022. High-resolution satellite products improve hydrological modeling in northern Italy. Hydrol. Earth Syst. Sci. 26 (14), 3921–3939. https://doi.org/10.5194/hess-26-3921-2022.

Allen, R.G., Tasumi, M., Trezza, R., 2007. Satellite-based energy balance for mapping evapotranspiration with internalized calibration (METRIC) – Model. ASCE J. Irr. Drain. Eng. 133, 380–394. https://doi.org/10.1061/(ASCE)0733-9437(2007)133:4(380).

Allen, R.G., Pereira, L.S., Howell, T.A., Jensen, M.E., 2011. Evapotranspiration information reporting: I. Factors governing measurement accuracy. Agr. Water Manag. 98, 899–920. https://doi.org/10.1016/j.agwat.2010.12.015.

Anderson, M.C., Norman, J.M., Diak, G.R., Kustas, W.P., Mecikalski, J.R., 1997. A two-source time-integrated model for estimating surface fluxes using thermal infrared remote sensing. Remote Sens. Environ. 60 (2), 195–216. https://doi.org/10.1016/S0034-4257(96)00215-5.

Anderson, M.C., Norman, J.M., Mecikalski, J.R., Torn, R.D., Kustas, W.P., Basara, J.B., 2004. A multi-scale remote sensing model for disaggregating regional fluxes to micrometeorological scales. J. Hydrometeorol. 5, 343–363. https://doi.org/10.1175/1525-7541(2004)005<0343:AMRSMF>2.0.CO;2.

Anderson, M.C., Norman, J.M., Mecikalski, J.R., Otkin, J.A., Kustas, W.P., 2007. A climatological study of evapotranspiration and moisture stress across the continental U.S. based on thermal remote sensing: I. Model formulation. J. Geophys. Res. 112, D10117. https://doi.org/10.1029/2006JD007506.

Anderson, M.C., Hain, C.R., Wardlow, B., Mecikalski, J.R., Kustas, W.P., 2011. Evaluation of a drought index based on thermal remote sensing of evapotranspiration over the continental U.S. J. Climate 24, 2025–2044. https://doi.org/10.1175/2010JCLI3812.1.

Anderson, M.C., Gao, F., Knipper, K., Hain, C., Dulaney, W., Baldocchi, D.D., Eichelmann, E., Hemes, K.S., Yang, Y., Medellín-Azuara, J., Kustas, W.P., 2018. Field-scale assessment of land and water use change over the California Delta using remote sensing. Remote Sens. 10, 889. https://doi.org/10.3390/rs10060889.

Arsenault, R., Gatien, P., Renaud, B., Brissette, F., Martel, J.L., 2015. A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation. J. Hydrol. 529, 754–767. https://doi.org/10.1016/j.jhydrol.2015.09.001.

Autovino, D., Minacapilli, M., Provenzano, G., 2016. Modelling bulk surface resistance by MODIS data and assessment of MOD16A2 evapotranspiration product in an irrigation district of Southern Italy. Agr. Water Manag. 167, 86–94. https://doi.org/10.1016/j.agwat.2016.01.006.

Bai, P., 2023. Comparison of remote sensing evapotranspiration models: Consistency, merits, and pitfalls. J. Hydrol. 617 (A), 128856. https://doi.org/10.1016/j.jhydrol.2022.128856.

Balsamo, G., Beljaars, A., Scipal, K., Viterbo, P., van den Hunk, B., Hirschi, M., Betts, A. K., 2009. A revised hydrology for the ECMWF model: Verification from field site to terrestrial water storage and impact in the integrated forecast system. J. Hydrometeorol. 10 (3), 623–643. https://doi.org/10.1175/2008JHM1068.1.

Bastiaanssen, W.G.M., Menenti, M., Feddes, R.A., Holtslag, A.A.M., 1998. The surface energy balance algorithm for land (SEBAL): Part 1 formulation. J. Hydrol. 212–213, 198–212. https://doi.org/10.1016/S0022-1694(98)00253-4.

Beck, H.E., van Dijk, A.I.J.M., de Roo, A., Dutra, E., Fink, G., Orth, R., Schellekens, J., 2017. Global evaluation of runoff from 10 state-of-the-art hydrological models. Hydrol. Earth Syst. Sci. 21, 2881–2903. https://doi.org/10.5194/hess-21-2881-2017.

Bhattarai, N., Shaw, S.B., Quackenbush, L.J., Im, J., Niraula, R., 2016. Evaluating five remote sensing based single-source surface energy balance models for estimating daily evapotranspiration in a humid subtropical climate. Int. J. Appl. Earth Obs. Geoinf. 49, 75–86. https://doi.org/10.1016/j.jag.2016.01.010.

Braca, G., Bussettini, M., Lastoria, B., Mariani, S., Piva, F., 2021. Il bilancio idrologico GIS based a scala nazionale su griglia regolare BIG BANG: metodologie e stime. Rapporto sulla disponibilità naturale della risorsa. In: Istituto Superiore per La Protezione e La Ricerca Ambientale, p. pp. [last access: September 2023].

Braca, G., Bussettini, M., Lastoria, B., Mariani, S., Piva, F., 2023. BIGBANG version 6.0 model processing. Italian Institute for Environmental Protection and Research – [last access: July 2023].

Branzei, R., Tijs, S., Timmer, J., 2000. Collecting information to improve decision-making. Int. Game Theory Rev. 3 (01), 1–12. https://doi.org/10.2139/ssrn.247773.

Cammalleri, C., Rallo, G., Agnese, C., Ciraolo, G., Minacapilli, M., Provenzano, G., 2013. Combined use of eddy covariance and sap flow techniques for partition of ET fluxes and water stress assessment in an irrigated olive orchard. Agr. Water Manag. 120, 89–97. https://doi.org/10.1016/j.agwat.2012.10.003.

Cammalleri, C., Micale, F., Vogt, J.V., 2015. On the value of combining different modelled soil moisture products for European drought monitoring. J. Hydrol. 525, 547–558. https://doi.org/10.1016/j.jhydrol.2015.04.021.

Cammalleri, C., Vogt, J.V., Bisselink, B., de Roo, A., 2017. Comparing soil moisture anomalies from multiple independent sources over different regions across the globe. Hydrol. Earth Syst. Sci. 21, 6329–6343. https://doi.org/10.5194/hess-21-6329-2017.

Castelli, M., 2021. Evapotranspiration changes over the European Alps: Consistency of trends and their drivers between the MOD16 and SSEBop algorithms. Remote Sens. 13 (21), 4316. https://doi.org/10.3390/rs13214316.

Chen, F., Crow, W.T., Ciabatta, L., Filippucci, P., Panegrossi, G., Marra, A.C., Puca, S., Massari, C., 2021. Enhanced large-scale validation of satellite-based land rainfall products. J. Hydrometeorol. 22 (2), 245–257. https://doi.org/10.1175/JHM-D-20-0056.1.

Copernicus Land Monitoring Service (CLMS), 2021. CORINE Land Cover: User Manual v1.0. European Environ.ment Agency (EEA), 129 pp. Available at: https://land.copernicus.eu/user-corner/technical-library/clc-product-user-manual [last access: July 2023].

Courault, D., Seguin, B., Olioso, A., 2005. Review on estimation of evapotranspiration from remote sensing data: From empirical to numerical modeling approaches. Irr. Drain. Syst. 19, 223–249. https://doi.org/10.1007/s10795-005-5186-0.c.

de Roo, A., Wesseling, C., van Deusen, W., 2000. Physically based river basin modelling within a GIS: the LISFLOOD model. Hydrol. Process. 14, 1981–1992. https://doi.org/10.1002/1099-1085(20000815/30)14:11/12<1981::AID-HYP49>3.0.CO;2-F.

De Santis, D., D'Amato, C., Bartkowiak, P., Azimi, S., Castelli, M., Rigon, R., Massari, C., 2022. Evaluation of remotely-sensed evapotranspiration datasets at different spatial and temporal scales at forest and grassland sites in Italy. 2022 IEEE Workshop on Metrology for Agriculture and Forestry (MetroAgriFor), Perugia, Italy, 356-361. doi: 10.1109/MetroAgriFor55389.2022.9964755.

DeFries, R.S., Hansen, M., 2010. ISLSCP II University of Maryland Global Land Cover Classification 1992-1993. In: Hall, Forrest G., G. Collatz, B. Meeson, S. Los, E. Brown de Colstoun, and D. Landis (eds.). ISLSCP Initiative II Collection. Data set. Available online [http://daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA. doi:10.3334/ORNLDAAC/969.

Di Giovanni, A., Di Curzio, D., Pantanella, D., Picchi, C., Rusi, S., 2023. Estimating a reliable water budget at a basin scale: A comparison between the geostatistical and traditional methods (Foro River basin, central Italy). Water 15 (23), 4083. https://doi.org/10.3390/w15234083.

Dong, J., Lei, F., Wei, L., 2020. Triple collocation based multi-source precipitation merging. Front. Water 2–2020. https://doi.org/10.3389/frwa.2020.00001.

Draper, C., Reichle, R., de Jeu, R., Naeimi, V., Parinussa, R., Wagner, W., 2013. Estimating root mean square errors in remotely sensed soil moisture over continental scale domains. Remote Sens. Environ. 137, 288–298. https://doi.org/10.1016/j.rse.2013.06.013.

Duan, Z., Duggan, E., Chen, C., Gao, H., Dong, J., Liu, J., 2021. Comparison of traditional method and triple collocation analysis for evaluation of multiple gridded precipitation products across Germany. J. Hydrometeorol. 22 (11), 2983–2999. https://doi.org/10.1175/JHM-D-21-0049.1.

Eurostat, 2022. Statistical regions in the European Union and partner countries: NUTS and statistical regions 2021. 2022 edition, Luxembourg: Publications Office of the European Union, 188 pp. doi:10.2785/321792.

Faroux, S., Kaptue Tchuent, A.T., Roujean, J.-L., Masson, V., Martin, E., Le Moigne, P., 2013. ECOCLIMAP-II/Europe: a twofold database of ecosystems and surface parameters at 1 km resolution based on satellite information for use in land surface, meteorological and climate models. Geosci. Model Dev. 6, 563–582. https://doi.org/10.5194/gmd-6-563-2013.

Ghilain, N., Arboleda, A., Gellens-Meulenberghs, F., 2011. Evapotranspiration modelling at large scale using near-real time MSG SEVIRI derived data. Hydrol. Earth Syst. Sci. 15, 771–786. https://doi.org/10.5194/hess-15-771-2011.

Gruber, A., Su, C.-H., Zwieback, S., Crow, W., Dorigo, W., Wagner, W., 2016. Recent advances in (soil moisture) triple collocation analysis. Int. J. Appl. Earth Obs. 45, 200–211. https://doi.org/10.1016/j.jag.2015.09.002.

Hain, C.R., Anderson, M.C., 2017. Estimating morning change in land surface temperature from MODIS day/night observations: Applications for surface energy balance modelling. Geophys. Res. Lett. 44, 9723–9733. https://doi.org/10.1002/2017GL074952.

Hassler, B., Lauer, A., 2021. Comparison of reanalysis and observational precipitation datasets including ERA5 and WFDE5. Atmos. 12 (11), 1462. https://doi.org/10.3390/atmos12111462.

Hu, T., Mallick, K., Hitzelberger, P., Didry, Y., Boulet, G., Szantoi, Z., Koetz, B., Alonso, I., Pascolini-Campbell, M., Halverson, G., Cawse-Nicholson, K., Hulley, G.C., Hook, S., Bhattarai, N., Olioso, A., Roujean, J.-L., Gamet, P., Su, B., 2023. Evaluating European ECOSTRESS hub evapotranspiration products across a range of soil-atmospheric aridity and biomes over Europe. Water Resour. Res. 59(8), e2022WR034132. doi: 10.1029/2022WR034132.

Hu, G., Jia, L., Menenti, M., 2015. Comparison of MOD16 and LSA-SAF MSG evapotranspiration products over Europe for 2011. Remote Sens. Environ. 156, 510–526. https://doi.org/10.1016/j.rse.2014.10.017.

Kim, H.W., Hwang, K., Mu, Q., Lee, S.O., Choi, M., 2012. Validation of MODIS 16 global terrestrial evapotranspiration products in various climates and land cover types in Asia. KSCE J. Civil Eng. 16, 229–238. https://doi.org/10.1007/s12205-012-0006-1.

Kirtman, B.P., Min, D., Infanti, J.M., Kinter, J.L., Paolino, D.A., Zhang, Q., Van Den Dool, H., Saha, S., Mendez, M.P., Becker, E., Peng, P., Tripp, P., Huang, J., DeWitt, D. G., Tippett, M.K., Barnston, A.G., Li, S., Rosati, A., Schubert, S.D., Rienecker, M., Suarez, M., Li, Z.E., Marshak, J., Lim, Y.-K., Tribbia, J., Pegion, K., Merryfield, W.J., Denis, B., Wood, E.F., 2014. The North American multimodel ensemble: phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. Bul. Am. Meteorol. Soc. 95 (4), 585–601. https://doi.org/10.1175/BAMS-D-12-00050.1.

Li, X., Zhang, W., Vermeulen, A., Dong, J., Duan, Z., 2023. Triple collocation-based merging of multi-source gridded evapotranspiration data in the Nordic Region. Agr. Forest Meteorol. 335, 109451. https://doi.org/10.1016/j.agrformet.2023.109451.

LSA SAF, 2016. Algorithm Theoretical Basis Document (ATBD) for Daily Evapotranspiration (DMET v2). SAF/LAND/RMIB/ATBD_METv2_DMET. Available on-line at http://landsaf.meteo.pt [last access: July 2023].

LSA SAF, 2023. Validation Report (VR) for Evapotranspiration and Surface Fluxes. SAF/LAND/RMI/VR/2.0. Available online at http://landsaf.meteo.pt [last access: June 2024].

McColl, K.A., Vogelzang, J., Konings, A.G., Entekhabi, D., Piles, M., Stoffelen, A., 2014. Extended Triple Collocation: Estimating errors and correlation coefficients with respect to an unknown target. Geophys. Res. Lett. 41, 6229–6236. https://doi.org/10.1002/2014GL061322.

McKinney, D.C., Cai, X., Rosegrant, M.W., Ringler, C., Scott, C.A., 1999. Modeling Water Resources Management at the Basin Level: Review and Future Directions ix, 59p. SWIM paper 6.

Melo, D.C.D., Anache, J.A.A., Borges, V.P., Miralles, D.G., Martens, B., Fisher, J.B., et al., 2021. Are remote sensing evapotranspiration models reliable across South American ecoregions? Water Resour. Res. 57, e2020WR028752. doi:org/10.1029/2020WR028752.

Melton, F.S., Huntington, J., Grimm, R., Herring, J., Hall, M., Rollison, D., Erickson, T., Allen, R., Anderson, M., Fisher, J.B., Kilic, A., Senay, G.B., Volk, J., Hain, C., Johnson, L., Ruhoff, A., Blankenau, P., Bromley, M., Carrara, W., Daudert, B., Doherty, C., Dunkerly, C., Friedrichs, M., Guzman, A., Halverson, G., Hansen, J., Harding, J., Kang, Y., Ketchum, D., Minor, B., Morton, C., Ortega-Salazar, S., Ott, T., Ozdogan, M., ReVelle, P.M., Schull, M., Wang, C., Yang, Y., Anderson, R.G., 2022. OpenET: Filling a critical data gap in water management for the western United States. J. Am. Water Resour. Ass. 58 (6), 971–994. https://doi.org/10.1111/1752-1688.12956.

Miralles, D.G., Crow, W.T., Cosh, M.H., 2010. Estimating spatial sampling errors in coarse-scale soil moisture estimates derived from point-scale observations. J. Hydrometeorol. 11 (6), 1423–1429. https://doi.org/10.1175/2010JHM1285.1.

Miralles, D.G., Holmes, T.R.H., De Jeu, R.A.M., Gash, J.H., Meesters, A.G.C.A., Dolman, A.J., 2011. Global land-surface evaporation estimated from satellite-based observations. Hydrol. Earth Syst. Sci. 15, 453–469. https://doi.org/10.5194/hess-15-453-2011.

Monteith, J.L., 1965. Evaporation and environment. In Symposia of the Society for Experimental Biology 19, 205–234.

Mu, Q., Zhao, M., Running, S.W., 2011. Improvements to a MODIS global terrestrial evapotranspiration algorithm. Remote Sens. Environ. 115 (8), 1781–1800. https://doi.org/10.1016/j.rse.2011.02.019.

Norman, J.M., Kustas, W.P., Humes, K.S., 1995. Source approach for estimating soil and vegetation energy fluxes in observations of directional radiometric surface temperature. Agric. Water Meteorol. 77 (3–4), 263–293. https://doi.org/10.1016/0168-1923(95)02265-Y.

Pan, M., Fisher, C.K., Chaney, N.W., Zhan, W., Crow, W.T., Aires, F., Entekhabi, D., Wood, E.F., 2015. Triple collocation: Beyond three estimates and separation of structural/non-structural errors. Remote Sens. Environ. 171, 299–310. https://doi.org/10.1016/j.rse.2015.10.028.

Pan, S., Pan, N., Tian, H., Friedlingstein, P., Sitch, S., Shi, H., Arora, V.K., Haverd, V., Jain, A.K., Kato, E., Lienert, S., Lombardozzi, D., Nabel, J.E.M.S., Ottlcé, C., Poulter, B., Zaehle, S., Running, S.W., 2020. Evaluation of global terrestrial evapotranspiration using state-of-the-art approaches in remote sensing, machine learning and land surface modeling. Hydrol. Earth Syst. Sci. 24, 1485–1509. https://doi.org/10.5194/hess-24-1485-2020.

Panagos, P., Van Liedekerke, M., Jones, A., Montanarella, L., 2012. European Soil Data Centre: Response to European policy support and public data requirements. Land Use Policy 29 (2), 329–338. https://doi.org/10.1016/j.landusepol.2011.07.003.

Park, J., Baik, J., Choi, M., 2023. Triple collocation-based multi-source evaporation and transpiration merging. Agr. Forest Meteorol. 331, 109353. https://doi.org/10.1016/j.agrformet.2023.109353.

Pastorello, G., Trotta, C., Canfora, E., et al., 2020. The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. Sci Data 7, 225. https://doi.org/10.1038/s41597-020-0534-3.

Pinna, M., 1970. Contributo alla classificazione del clima d'Italia. Rivista Geografica Italiana 2, 129–152.

Ramoelo, A., Majozi, N., Mathieu, R., Jovanovic, N., Nickless, A., Dzikiti, S., 2014. Validation of global evapotranspiration product (MOD16) using flux tower data in the African savanna. South Africa. Remote Sens. 6 (8), 7406–7423. https://doi.org/10.3390/rs6087406.

Rana, G., Katerji, N., 2000. Measurement and estimation of actual evapotranspiration in the field under Mediterranean climate: a review. Eur. J. Agron. 13 (2–3), 125–153. https://doi.org/10.1016/S1161-0301(00)00070-8.

Rossi, L., Naumann, G., Gabellani, S., Cammalleri, C., 2023. A combined index to characterize agricultural drought in Italy at municipality scale. J. Hydrol. Reg. Stud. 47, 101404. https://doi.org/10.1016/j.ejrh.2023.101404.

Ruggieri, G., Allocca, V., Borfecchia, F., Cusano, D., Marsiglia, P., De Vita, P., 2021. Testing evapotranspiration estimates based on MODIS satellite data in the assessment of the groundwater recharge of karst aquifers in southern Italy. Water 13, 118. https://doi.org/10.3390/w13020118.

Running, S., Mu, Q., Zhao, M., 2017. MOD16A2 MODIS/Terra Net Evapotranspiration 8-Day L4 Global 500m SIN Grid V006. Distributed by NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/MODIS/MOD16A2.006.

Saha, S., Moorthi, S., Pan, H.-L., Wu, X., et al., 2010. The NCEP climate forecast system reanalysis. Bull. Amer. Meteorol. Soc. 91, 1015–1057. https://doi.org/10.1175/2010BAMS3001.1.

Salvati, L., Petitta, M., Ceccarelli, T., Perini, L., Di Battista, F., Venezian Scarascia, M.E., 2008. Italy's renewable water resources as estimated on the basis of the monthly water balance. Irr. Drain. 57 (5), 507–515. https://doi.org/10.1002/ird.380.

Senay, G.B., Budde, M., Verdin, J.P., 2011. Enhancing the Simplified Surface Energy Balance (SSEB) approach for estimating landscape ET: Validation with the METRIC model. Agr. Water Manag. 98, 606–618. https://doi.org/10.1016/j.agwat.2010.10.014.

Senay, G.B., Bohms, S., Singh, R.K., Gowda, P.H., Velpuri, N.M., Alemu, H., Verdin, J.P., 2013. Operational evapotranspiration mapping using remote sensing and weather datasets: A new parameterization for the SSEB approach. J. Am. Water Resour. Ass. 49 (3), 577–591. https://doi.org/10.1111/jawr.12057.

Senay, G.B., Kagone, S., Velpuri, N.M., 2020. Operational global actual evapotranspiration: Development, evaluation, and dissemination. Sensors 20 (7), 1915. https://doi.org/10.3390/s20071915.

Singh, S.K., Pahlow, M., Duan, Q., Griffiths, G., 2019. Improving hydrological multi-model prediction by elimination of double counting in the ensemble. J. Hydrol. (NZ) 58 (2), 81–103. https://www.jstor.org/stable/26912150.

Stoffelen, A., 1998. Towards the true near-surface wind speed: Error modeling and calibration using triple collocation. J. Geophys. Res. 103 (C4), 7755–7766. https://doi.org/10.1029/97JC03180.

Sun, H., Yong, Y., Wu, R., Gui, D., Xue, J., Liu, Y., Yan, D., 2019. Improving estimation of cropland evapotranspiration by the Bayesian model averaging method with surface energy balance models. Atmos. 10 (4), 188. https://doi.org/10.3390/atmos10040188.

Thompson, P.D., 1977. How to improve accuracy by combining independent forecasts. Month. Weather Rev. 105 (2), 228–229. https://doi.org/10.1175/1520-0493(1977)105<0228:HTIABC>2.0.CO;2.

Twine, T.E., Kustas, W.P., Norman, J.M., Cook, D.R., Houser, P.R., Meyers, T.P., Prueger, J.H., Starks, P.J., Wesely, M.L., 2000. Correcting eddy-covariance flux underestimates over a grassland. Agr. Forest Meteorol. 103, 279–300. https://doi.org/10.1016/S0168-1923(00)00123-4.

van Der Knijff, J.M., Younis, J., De Roo, A.P.J., 2010. LISFLOOD: a GISbased distributed model for river basin scale water balance and flood simulation, Int. J. Geograph. Inf. Sci. 24(2), 189–212. doi:10.1080/ 13658810802549154.

Vinukollu, R.K., Meynadier, R., Sheffield, J., Wood, E.F., 2011. Multi-model, multi-sensor estimates of global evapotranspiration: climatology, uncertainties and trends. Hydrol. Process. 25 (26), 3993–4010. https://doi.org/10.1002/hyp.8393.

Vogelzang, J., Stoffelen, A., 2022. On the accuracy and consistency of quintuple collocation analysis of in situ scatterometer and NWP winds. Remote Sens. 14 (18), 4552. https://doi.org/10.3390/rs14184552.

Volk, J.M., Huntington, J.L., Melton, F.S., Allen, R., Anderson, M., Fisher, J.B., Kilic, A., Ruhoff, A., Senay, G.B., Minor, B., Morton, C., Ott, T., Johnson, L., Comini de Andrade, B., Carrara, W., Doherty, C.T., Dunkerly, C.W., Friedrichs, M., Guzman, A., Hain, C., Halverson, G., Kang, Y., Knipper, K., Laipelt, L., Ortega-Salazar, S., Pearson, C., Parrish, G.E., Purdy, A., ReVelle, P., Wang, T., Yang, Y., 2024. Assessing the accuracy of OpenET satellite-based evapotranspiration data to support water

resource and land management applications. Nature Water 2, 193–205. https://doi.org/10.1038/s44221-023-00181-7.

Wang, K., Dickinson, R.E., 2012. A review of global terrestrial evapotranspiration: Observation, modeling, climatology, and climatic variability. Rev. Geophys. 50 (2), RG2005. https://doi.org/10.1029/2011RG000373.

Wang, S., Pan, M., Mu, Q., Shi, X., Mao, J., Brümmer, C., Jassal, R.S., Krishnan, P., Li, J., Black, T.A., 2015. Comparing evapotranspiration from eddy covariance measurements, water budgets, remote sensing, and land surface models over Canada. J. Hydrometeorol. 16, 1540–1560. https://doi.org/10.1175/JHM-D-14-0189.1.

Wanniarachchi, S., Sarukkalige, R., 2022. A review on evapotranspiration estimation in agricultural water management: Past, present, and future. Hydrol. 9 (7), 123. https://doi.org/10.3390/hydrology9070123.

World Meteorological Organization (WMO), 2017. WMO Guidelines on the Calculation of Climate Normals. WMO-No. 1203, Geneva, Switzerland. 29 pp. Available at: https://library.wmo.int/doc_num.php?explnum_id=4166 [Last access: July 2023].

Wood, E.F., Schubert, S.D., Wood, A.W., Peters-Lidard, C.D., Mo, K.C., Mariotti, A., Pulwarty, R.S., 2015. Prospects for advancing drought understanding, monitoring, and prediction. J. Hydrometeorol. 16 (4), 1636–1657. https://doi.org/10.1175/JHM-D-14-0164.1.

Xie, Q., Jia, L., Menenti, M., Hu, G., 2022. Global soil moisture data fusion by triple collocation analysis from 2011 to 2018. Sci. Data 9, 687. https://doi.org/10.1038/s41597-022-01772-x.

Yang, X.Q., Yong, B., Ren, L.L., Zhang, Y.Q., Long, D., 2017. Multi-scale validation of GLEAM evapotranspiration products over China via ChinaFLUX ET measurements. Int. J. Remote Sens. 38(20), 5688-5709. doi:10.1080/ 01431161.2017.1346400.

Yilmaz, M.T., Crow, W.T., 2014. Evaluation of assumptions in soil moisture triple collocation analysis. J. Hydrometeorol. 15 (3), 1293–1302. https://doi.org/10.1175/JHM-D-13-0158.1.

Yilmaz, M.T., Crow, W.T., Anderson, M.C., Hain, C., 2012. An objective methodology for merging satellite- and model-based soil moisture products. Water Resour. Res. 48, W11502. https://doi.org/10.1029/2011WR011682.

Zhang, B., Xia, Y., Long, B., Hobbins, M., Zhao, X., Hain, C., Li, Y., Anderson, M.C., 2020. Evaluation and comparison of multiple evapotranspiration data models over the contiguous United States: Implications for the next phase of NLDAS (NLDAS-Testbed) development. Agr. Forest Meteorol. 280, 107810. https://doi.org/10.1016/j.agrformet.2019.107810.

Zhu, W.B., Tian, S.R., Wei, J.X., Jia, S.F., Song, Z.K., 2022. Multi-scale evaluation of global evapotranspiration products derived from remote sensing images: Accuracy and uncertainty. J. Hydrol. 611, 127982. https://doi.org/10.1016/j.jhydrol.2022.127982.

Zwieback, S., Scipal, K., Dorigo, W., Wagner, W., 2012. Structural and statistical properties of the collocation technique for error characterization. Nonlin. Processes Gephys. 19, 69–80. https://doi.org/10.5194/npg-19-69-2012.