Revision of the paper HESS-2017-196

We would like to thank to both the anonymous reviewers for the thoughtful comments provided to our manuscript. A point-to-point reply to major and minor comments is reported below, as well as indication of the changes made to the text to address such comments.

Reply to Referee #1 Comments

General

This study investigates the potential of one model-based and two remotely sensed datasets for their use in an established global drought monitoring system. The study uses soil moisture from the Lisflood model, surface soil moisture from the microwave-based ESA CCI soil moisture dataset and land Surface temperature from MODIS. Random errors of these datasets are characterized with the triple collocation analysis (TCA), a firmly established method in soil moisture analysis. As such, apart from applying the TCA to a different triple of datasets, the study is not very innovative. However, since the datasets are expected to be used in an operational drought monitoring system, the results of this study are expected to have a large practical impact. This study is performed in a scientifically sound and clearly structured way and provide some interesting insights in the skill of the datasets considered. I therefore recommend its publication after addressing the following issues.

Thanks for your comment. Regarding the innovative aspect of the work, we agree that the application of triple collocation is nothing new in the scientific literature of soil moisture; however, there are very few studies analyzing modelled, microwave and thermal data, and none of them are focusing on the study areas reported here. Additionally, the use of standardized anomalies in TCA is uncommon in the literature even if it is a key aspect in drought studies. As you stated, the final goal of the study is to provide insight on the spatial distribution of the errors in our global drought observatory, hence the analysis of these three specific datasets (previously uninvestigated in TCA literature) and anomaly values is a novelty and a key requirement of our study. We clarified these novelties of the study in the new introduction section of the paper.

Major comments

1. The analyses in this study are based on monthly anomalies. Although theoretically it is feasible to do so, I wonder what the practical relevance of these results are. All datasets used are also available at a daily time step and already now the European Drought Observatory works with ten-day periods (dekades). Hence, also the error structures should be known at these time scales.

The use of time-aggregated data is a common practice in drought analyses in order to ensure the statistical robustness of the computed anomalies. Daily values are often too noisy to allow for a robust statistical analysis. The use of a monthly temporal scale is quite common in the drought literature, and it has been preferred (at the moment) over the ten-day scale (adopted in EDO) for operational/technical reasons. Indeed, the monthly time scale will be the one implemented in the GDO system as a first approach. Higher temporal resolutions will be tested and implemented in the future. We clarified this point in the revised version of the manuscript.

2. I was expecting a more thorough analysis on the consequences of using three datasets that represent very different layer depths, i.e.:

* ESA CCI soil moisture represents the upper 2 cm, * LST represents the skin temperature, which is driven both by surface soil moisture (influencing bare soil temperature) and root zone soil moisture (impacting vegetation canopy temperature) * LIS represents root zone soil moisture (but layer depth is not provided in the manuscript).

Typically, deeper and thicker layers have lower random errors than observations of the surface layer, but this also depends on the time scale you're looking at (e.g. daily observations typically have larger random errors than monthly averages). Is there a way you can test the impact of using such different layer depths, e.g. by using the surface layer of LISFLOOD?

We agree that the different vertical (as well as horizontal) resolution of the three datasets may lead to some discrepancies in the analysis. For this reason, we elaborated the data in order to try to minimize the possible discrepancies among the datasets. For instance, the use of monthly aggregated data is one of the techniques adopted to ensure minimizing the discrepancies between the three datasets related to the slower response of deeper soil layers (as you highlighted). Also, the use of standardized quantities allows reducing the effects of possible biases between soil moisture modelled at different depths.

We modified the text to highlight further the expedients adopted to try to minimize those issues, as well as highlighting where the different resolutions may lead to a better performance of one dataset over the others.

3. Some important information is missing (or not clearly provided) which is needed for a correct interpretation of the results: in which units are the TCA results expressed? Is it the fractional RMSE? If not, are the errors of each dataset expressed in its own data space or in a common data space provided by one of the models?

We better clarified that the errors are provided as dimensionless quantities (multiple of standard deviation), due to the standardization applied to the three datasets before performing the TCA. The standardization procedure also allows having all the three datasets in the same range of variability (zero mean, unitary standard deviation), which removes the need of a reference data space. Additionally, as briefly stated in the "Methods" section, we adopted the TCA "covariance notation", which does not require a common (arbitrary) reference dataset. We highlighted that information in the revised text.

Minor Comments

Line 16: why do you use the word proxy here? Only LST can be considered a proxy of soil moisture, while both LIS and ESA CCI are real estimates of soil moisture.

We rephrased this sentence.

Line 20: The official name is ESA CCI Soil Moisture or ESA CCI SM.

We reworded this definition of the product.

line 43: When agricultural droughts start affecting human welfare, people commonly use the term socioeconomic drought.

In our opinion, socioeconomic drought is a much larger concept that includes also other economic factors and that is separated from the classical "physical" classification meteorological/hydrological/agricultural. We prefer to leave this sentence as it is.

Lines 47/48: include references or URLs to these drought monitoring systems.

Done.

Lines 51/52: these citations refer to the formulation of drought indices rather than soil moisture.

We clarified that we were referring to soil moisture modelling in the context of drought.

Lines 83-97: please provide references to the various datasets discussed here (MODIS LST, ESA CCI SM, LISFLOOD).

References are reported in the successive sub-sections dedicated to each product. We prefer to leave this part of the introduction easy to read by avoiding further references.

Regarding the skill of LST-based soil moisture versus ESA CCI SM the authors should also refer to the ALEXI-based work of Fang et al. 2016 (http://www.sciencedirect.com/science/article/pii/S0303243415300404).

Thanks. We added some comments on this work in the results section.

Line 93: ESA CCI SM will soon be updated in NRT in the framework of Copernicus Climate Change Services (http://www.sciencedirect.com/science/article/pii/S0303243415300404).

Yes, we added this information to the text.

Line 97: The use of the term "models" is confusing and only applies to Lisflood. Replace "models" with "datasets" or "products", also throughout the rest of the manuscript.

Done.

Line 120: No definition is given of the root zone soil moisture simulated with Lisflood. What soil column is sampled?

Information have been added to sub-section 3.1.

line 128: I don't see how Pearson R would give information about the slope and biases between two datasets. Do you confuse it with the "regression function" between the datasets?

This is true only for the specific case of a regression analyses between two standardized quantities with zero mean and a unitary variance. The text has been amended to clarify this point.

Line 130: to my knowledge the correct name for this test is Student's t-test.

Done.

Line 148: spelling error: where -> were.

Done.

Line 168: it is unclear why you don't use the surface layer, which is much closer to the other two data sets. I suggest to repeat the TCA with the surface layer as well to see what impact is on the estimated errors of all datasets.

We agree that the LIS surface layer is closer to CCI but not necessarily to LST (which depth depends on vegetation coverage). We preferred root zone because it is more relevant for agricultural drought studies. Currently, we are testing the value of extrapolated "root-zone-like" soil moisture from skin CCI values, but this analysis is going beyond the goal of this paper.

Line 204: It may be worth checking whether using Microwave-based LSTs would lead to similar results (See Holmes et al., 2009; http://onlinelibrary.wiley.com/doi/10.1029/2008JD010257).

We are aware of the studies on microwave LST, but we do not think that these are relevant in this specific case study, since they compare quite well with MODIS data under clear sky conditions. The main advance of microwave-based LST dataset is to remove the limitation of thermal data during cloudy days, which is not impactful at monthly time scale. Also, microwave data have, at the moment, a coarser spatial resolution.

Line 225: provide version number of "current" version.

Done.

Line 233: Please check Terms and Conditions (http://www.esa-soilmoisture-cci.org/dataregistration/terms-and-conditions) for a correct citation of the data.

References have been updated.

Line 236-237: Only for the integration of SSM/I into the merged products the soil moisture signal is decomposed into seasonality and anomalies (Liu et al. (2012)).

Yes, you are right. However, this is just a very brief summary of the procedure and we think that the readers can find a very clear and detailed description of the procedure (if they are interested) in the cited paper.

Line 253: it is not clear whether the linear correlation is computed from the original signal or from the anomalies. Since your TCA implementation is based on the anomalies, also the correlation computation should be based on these.

Yes, it is computed on the anomalies. We clarified this in the text.

Line 282-283: Is the stronger correlation between CCI and LST not expected, as they represent more closely related soil layers? This is something you could test by including also the surface layer of Lisflood in your analysis.

In the new version of the text we added more considerations related to the explored vertical depth of each dataset. Yes, indeed it is expected that LST performs more closely to CCI in low-coverage areas (like Australia) and more closely to LIS over more vegetated areas.

Line 307: On one hand -> On the one hand.

Done.

Line 331: the results of this manuscript cannot be directly compared to those of Pierdicca et al., since the latter applies the TCA to daily observations.

Yes, this is true. This is the reason why we highlighted only some qualitative analogies. We clarified this difference in the two studies in the new version of the text.

Line 397: For what is skin soil moisture more reliable? Do you mean the estimates themselves? The estimation of soil moisture from microwave remote sensing may have large uncertainties over dry areas (e.g. Hahn et al., 2017: http://ieeexplore.ieee.org/document/7815274/).

We agree that under some circumstances microwave products can have large uncertainties over dry sandy soil. We rephrased the sentence to clarify that the presence of vegetation generally tends to reduce the reliability of microwave estimates.

Line 406: There are several studies that combine various datasets with different error characteristics, e.g. Liu et al., 2011 (http://www.hydrol-earth-syst-sci.net/15/425/2011/hess-15-425-2011.html); Beck et al., 2017 (http://www.hydrol-earth-syst-sci.net/21/589/2017/hess-21-589-2017.html); Yilmaz et al; 2012 (http://onlinelibrary.wiley.com/doi/10.1029/2011WR011682/full). Is there something that you can learn from these studies for your application?

We cited most of the reported authors (even if not specifically these papers) in out text already. The majority of these (and other) merging procedures are loosely based on a weighted average approach, with weights derived from the error analysis procedure. Particularly, the approach proposed in Yilmaz et al. (2012) is the one that we are currently implementing in our operational system, of which this study is the error characterization step. We highlighted this in the final section of the new version of the manuscript and we added some maps showing the actual spatial distribution and frequency of the weighting factors, as also suggested by rev. #2.

Reply to Referee #2 Comments

In the paper titled "Comparing soil moisture anomalies from multiple independent sources over different regions across the globe" authors have investigated the anomaly components of three products and then compared the errors of the standardized products over five different regions. Overall, the manuscript is written well and appropriate to the journal Hydrology and Earth System Sciences. However, there are some parts still need improvement:

- There are other soil moisture inter-comparison studies performed before at global scale, including the ones that have already implemented TCA. First of all authors should explicitly justify the need of anomaly comparisons at large scales (i.e., not per formed before over these locations using anomalies?). It is not all that clear what additional benefit do readers get from this study compared to the earlier studies (i.e., the analysis performed here are not performed before?). What is the new thing? Datasets? Locations? Anomaly components investigated instead of entire datasets? The information given in the introductions should better be tied with the overall goal.

We modified most of the final part of the introduction in order to highlight the novelty and the motivation of this study, as also requested by Rev. #1. In summary, most of the past TC studies were focused on soil moisture data rather than anomalies, with the latter being a key variable for drought monitoring that can behave quite differently from soil moisture itself. The inclusion of thermal remote sensing datasets is quite rare in the TC literature, and no studies on global scale are available to our knowledge. It follows that, even if the methodology adopted can be considered "standard", the data analyzed here are unique in terms of both variable investigated and datasets adopted. Finally, the operational focus of this study (error analysis to be used in a future implementation in a near-real time monitoring system) requires analyzing specifically those datasets that can be actually used in the GDO monitoring system.

What is the vertical support of the soil moisture product comparison performed here considering modeled soil moisture reflect root-zone while the MODIS LST and ESA CCI soil moisture products reflect the top couple cm depths. How does it relate to the overall framework of the study (agricultural drought monitoring while root-zone soil moisture lies at the hearth of such analyses in general)? Surface skin soil moisture is a good indicator for agricultural drought?

As replied to the second major comment of rev. #1, we adopted some pre-processing procedures in order to minimize likely discrepancies between datasets with different vertical resolution. In this context, the choice of a monthly-aggregation period aimed at removing the time shift between data referring to different soil depths, whereas standardized anomalies removed biases among the three datasets. The obtained results seem to support the idea that these expedites were able to make the datasets comparable overall.

These considerations were further highlighted in the new version of the manuscript.

Percentage error variance information is not all that helpful, perhaps actual standard deviations (volumetric error for the model and satellite soil moisture products and K for the LST product) would be more helpful (i.e., how do these errors relate to specific mission goals of 4%). Or at least authors should justify why presentation of standardized error variance is a better thing to do compared to actual error variance.

Actual error variance (i.e., in terms of volumetric water content or degree) cannot be reported in this study, since standardized variables (normalized z-scores) are investigated for the above mentioned reasons. The use of standardized quantities is justified by the needs of drought monitoring, as clarified in the new version of the manuscript.

Error variance comparisons of three datasets in space is done, but it would be helpful if more is given. For example, specific pattern between the error variance and vegetation/precipitation/elevation distributions? Any one better under such and such conditions (instead of only locations)? Why better? A dedicated paragraph would be very helpful.

We agree that it would be helpful to be able to provide more insight on the spatial patterns of the errors. However, the obtained results do not provide clear evidence of the suggested relations, behind some general behaviors (LIS perform better over well-monitored agricultural areas whereas CCI perform better over remote dry areas). We highlighted the inability to infer further on spatial patterns on the new version of the discussion and conclusion sections.

Combination of different datasets is spelled out in the introduction (final paragraph where the goal of the study is stated) but not performed (it gives the impression that this study will merge different products; perhaps it should have given all the necessary inputs are available including the error sources of each product).

We agree that this sentence may be misleading, and we rephrased the text to clarify this point. Even if the final goal of our project is to provide an operational ensemble product for drought monitoring, the goal of this specific study is "limited" to the characterization of spatial errors for the three selected datasets. The output of this error analysis is used to obtain statistical robust weighting factors for a reliable ensemble product to be used in GDO, and maps of these weighting factors have been added to the new version of the manuscript. Indeed, a first version of this product is already available in a prototype form in GDO, but a full implementation of the ensemble (based on the outcome of this study) is subordinate to the future availability of CCI in near-real time.

In the new version of the manuscript we clarified that the ensemble product is not the final goal of the reported study, but that the goal of the research is to spatially characterize the errors to be used in the future within the operational system. However, more details on the weighting factor have been added to partially fulfil your request.

Some background discussion about the TCA dataset requirements/assumptions (e.g., length? See Zwieback et al, 2012, doi:10.5194/npg-19-69-2012).

We added these considerations to the methods section, as well as some relevant references.

L98: “two folds” L102: “to develop a suitable combination procedure for a near-real time detection of the occurrence of ecosystem drought events”.. More specifics. How this will be performed? Using TCA errors to calculate the weights in a merging algorithm?

Yes. As now clarified in the text, the merging procedure will be based on a weighted average with weights derived from the TC errors. Details on the spatial distribution of the weighting factors have been added; however, the implementation of such ensemble product is not yet fully developed in the operational GDO system.

L112: revise “in order to make directly comparable the different datasets”.

Done.

Comparing soil moisture anomalies from multiple independent sources over different regions across the globe

Carmelo Cammalleri, Jürgen V. Vogt, Bernard Bisselink, Ad de Roo

European Commission, Joint Research Centre (JRC), Ispra, Italy.

Correspondence to: C. Cammalleri, European Commission - Joint Research Centre, via E. Fermi 2749, I-21027 Ispra (VA), Italy. Bldg. ~~100~~26b, Room ~~1208~~140, TP ~~122~~267. Phone: +39 (0)332.78.9869, e-mail: carmelo.cammalleri@ec.europa.eu.

Abstract: Agricultural drought events can affect large regions across the World, implying the urge for a suitable global tool for an accurate monitoring of this phenomenon. Soil moisture anomalies are considered a good metric to capture the occurrence of agricultural drought events, and they have become an important component of several operational drought monitoring systems. In the framework of the JRC Global Drought Observatory (GDO, http://edo.jrc.ec.europa.eu/gdo/) the suitability of ~~modelled and/or satellite-derived proxy of soil moisture anomalies~~of a ~~was investigated. In this study,~~ three datasets ~~have been evaluated~~ as possible ~~proxies~~ representation of root zone soil moisture anomalies has been evaluated: (1) soil moisture from the Lisflood distributed hydrological model (namely LIS), (2) remotely sensed ~~land~~ Land ~~surface~~ Surface ~~temperature~~ Temperature data from the MODIS satellite (namely LST), and (3) the ESA Climate Change Initiative combined passive/active microwave skin soil moisture dataset ~~developed by ESA~~ (namely CCI). Due to the independency of these three datasets, the Triple Collocation (TC) technique has been applied, aiming at quantifying the likely error associated to each dataset in comparison to the unknown true status of the system. TC analysis was performed on five macro-regions (namely North America, Europe, India, Southern Africa and Australia) detected as suitable for the experiment, providing insight into the mutual relationship between these datasets as well as an assessment of the accuracy of each method. Even if no definitive statement on the spatial distribution of errors can be provided, ~~A~~a clear outcome of the TC analysis is

the good performance of remote sensing datasets, especially CCI, over dry regions such as Australia and Southern Africa, whereas the outputs of LIS seem to be more reliable over areas that are well monitored through meteorological ground station networks, such as North America and Europe. In a global drought monitoring system, the~~se~~ results of the error analysis ~~can be~~are used to design a weighted-average ensemble system that exploits the advantages of each dataset.

1. Introduction

Drought is a recurring natural extreme, triggered by lower than normal rainfall, often exacerbated by a strong evaporative demand due to high temperatures and strong winds. Drought events may occur in all climates and in most parts of the world, since drought is defined as a temporary deviation from the local normal condition. Due to the usually wide extension of the interested area, drought affects millions of people across the Globe each year (Wilhite, 2000).

On the basis of the economic and natural sectors impacted by this phenomenon, a drought event is usually classified in meteorological, agricultural and hydrological drought, depending on the persistence of the water deficit within the hydrological cycle. Of particular interest for this study are the agricultural (or ecosystem) drought events, defined as prolonged periods with drier than usual soils that negatively affect vegetation growth and crop production, and, as a consequence, human welfare (Dai, 2011).

Soil moisture is commonly seen as one of the most suitable variables to monitor and quantify the impact of water shortage on vegetated lands due to its effects on the terrestrial biosphere and the feedback into the atmospheric system, as highlighted by the inclusion of time-aggregated soil moisture anomalies (e.g., monthly) in numerous drought monitoring systems at regional to continental scales (i.e., European Drought Observatory, http://edo.jrc.ec.europa.eu; United States Drought Monitor, http://droughtmonitor.unl.edu; African Flood and Drought Monitor, http://hydrology.princeton.edu/adm/; among others).

In the context of drought monitoring, the ~~S~~soil moisture ~~monitoring~~ dynamic over large areas is usually ~~obtained~~ modelled through either distributed hydrological models or land-surface schemes of climate models (Crow et al., 2012; Sheffield et al., 2004), as well as by thermal or passive/active microwave remote sensing-derived quantities (see e.g., Anderson et al., 2007; Houborg et al., 2012; Mo et al., 2010). Particularly, Iinwith regards to a the context of a global-scale -drought monitoring systemmonitoring, remote sensing-based approaches have the advantage of an intrinsic worldwide coverage, but the drawbacks, in the case of microwave sensors, of exploring only the first few centimeters of soil and a decreasing sensitivity with the increase of vegetation coverage (Jackson, 2006). In the case of thermal data, the lack of coverage during cloudy conditions and the nontrivial connection between thermal and soil moisture signals (Price, 1980) are other limitations. On the contrary, diagnostic models allow for a continuous monitoring of soil moisture at the desired soil depths, but the accuracy of the data is constrained by uncertainties in the parameterization of soil hydrological characteristics, as well as by the actual availability of near-real time reliable meteorological forcing data. Generally, the use of in-situ observations for large area monitoring is limited, mainly due to the lack of long records, the sparseness of recording stations and the high spatial heterogeneity of soil moisture fields.

It follows that both satellite measurements and model predictions are subject to errors and uncertainties that need to be accounted for in their interpretation and application (Gruber et al., 2016). This also suggests that a monitoring system based on a single model is rarely capable to provide global reliable estimates, and a combination of different data sources is desirable in order to minimize the errors in the detection of drought events. Recently, Cammalleri et al. (2015) demonstrated the value of an ensemble of modelled soil moisture anomalies for drought monitoring over Europe, similarly to the findings of the U.S. National Land Data Assimilation System (NLDAS) (Dirmeyer et al., 2006). However, a key point in combining different modelled data is the need to estimate the affinity and divergence between the models across the modelling domain.

In the most recent years, the Triple Collocation (TC) technique (Stoffelen, 1998) has been established as a practical approach to evaluate the unknown error variance (with respect to the truth) of three mutually independent measurement systems without knowing the “true” status of the system (Yilmaz and Crow, 2014). This technique has been widely applied in hydrology to estimate errors in soil moisture, as well as to evaluate precipitation and vegetation property indicators (Dorigo et al., 2010; McColl et al., 2014). One key requirement in TC is the existence of linearity between the three estimates and the truth, which can fail in the case of strongly seasonal geophysical variables such as soil moisture (Su et al., 2014). Luckily, drought monitoring systems are usually based on soil moisture anomalies rather than actual values, hence providing a partial remedy to this problem and making soil moisture anomalies directly suitable for this methodology (Miralles et al., 2010). However, since most of TC studies focused on soil moisture dynamics rather than standardized anomalies, specific analyses are required to evaluate ~~spatially~~ the accuracy of each dataset across the spatial domain.

In the frame of an operational monitoring of agriculture and ecosystem drought, the availability of soil moisture, or proxy~~,~~ datasets available in near-real time is crucial; within the Global Drought Observatory (GDO, http://edo.jrc.ec.europa.eu/gdo/), developed by the Joint Research Centre (JRC) of the European Commission, the soil moisture outputs of the Lisflood hydrological model and the land surface temperature (LST) anomalies derived from the Moderate-Resolution Imaging Spectroradiometer (MODIS) onboard the Terra satellite have been detected as suitable datasets for a near-real time monitoring. In particular, Cammalleri and Vogt (2016) have highlighted how monthly-average LST anomalies represent the best proxy of soil moisture variations across different climates in Europe when compared to other LST-derived quantities.

As a third dataset for the TC analysis, the combined active/passive microwave soil moisture dataset produced by the European Space Agency (ESA) in the context of the Climate Change Initiative (CCI) is used; even if this dataset is not currently updated in near-real time, it represents a valuable reference dataset for a global consistent time-series of microwave-based soil moisture maps (also, near-real time updating is foreseen in the framework of the Copernicus Climate Change Services).

The agreement between~~of~~ anomaly time-series derived from these three products has~~ve~~ not been fully investigated in the literature, especially at global scale; hence, given the independency of the three sources of data ~~The use of three independent sources of data~~ (hydrological model, thermal and microwave remote sensing) and the likely fulfilling of the main TC key ~~helps ensuring to fulfill a key~~ hypothesis ~~of TC, which is the~~ (i.e., independency between the errors of the three ~~models~~datasets), the

TC approach seems suitable for quantifying the spatial distribution of the errors associated to each dataset.

Following these considerations, ~~T~~the overall goal of this study is twofold. First, the agreement between the monthly anomalies of the three ~~models~~ datasets is evaluated, in order to identify the macro-areas where a reliable monitoring of soil moisture extreme conditions can be performed according to these three ~~available~~ datasets available globally and suitable for use in a near-real time monitoring system. Second, the TC analysis is performed over those macro-areas to quantify the spatial distribution of the expected random errors for each model compared to the unknown true status. Ultimate objective of the error analysis reported in this study is to provide information on the accuracy of the datasets that can be ~~in order to~~ -injected into a weighted-average ~~develop a suitable combination~~ ensemble procedure for a near-real time detection of the occurrence of ecosystem drought events, contributing ~~. Both goals will contribute to~~to the future development of a robust agricultural drought monitoring index within the GDO system.

2. Methods

Drought events are commonly defined as prolonged periods during which a given drought indicator significantly deviates from the usual condition for the specific site and period (e.g., soil moisture content is lower than the climatology). Following this definition, this study will focus on standardized z-score values in order to make the different datasets directly comparable ~~the different datasets~~ (i.e., minimizing the differences related to seasonality, soil depth, etc.). Specifically, monthly z-score values, or anomalies, are evaluated as:

$$z_{x,i,k} = \frac{x_{i,k} - \mu_{x,i}}{\sigma_{x,i}} \quad (1)$$

where $x_{i,k}$ is the monthly average variable for the i-th month at the k-th year, $\mu_{x,i}$ and $\sigma_{x,i}$ are the long-term average and standard deviation of the variable x for the i-th month, respectively. The baseline period adopted to compute the twelve ~~reference~~ μ and σ ~~twelve~~ monthly reference values should be of

15-30 years in order to ensure a stable benchmark. The three datasets used here, as described in the next section, are the root zone soil moisture data from the Lisflood model (x = LIS), the ESA Climate Change Initiative skin soil moisture microwave combined product (x = CCI) and the thermal remote sensing derived L~~l~~and S~~s~~urface T~~t~~emperature (x = LST); in the case of LST data, the sign of the anomalies is reversed due to the expected inverse relationship between soil moisture and LST.

The monthly aggregation period is chosen to ensure a statistical robustness of the computed anomalies, as well as to minimize the presence of missing data in the remote sensing datasets due to sub-optimal acquisition conditions (e.g., cloudy days for LST). The transition from daily data to monthly aggregated values also ensures a reduction in the likely discrepancies among the three datasets introduced by the differences in the explored soil depth, since the phase shift in time-aggregated quantities is usually less marked (Campbell and Norman, 1998). Additionally, the ~~time-series of~~ anomalies computed according to Eq. (1), ~~are~~ characterized by a null average and a unitary standard deviation, ~~making~~ allow for a ~~a~~ direct comparison of the different datasets ~~simpler~~ thanks to the removal of potential biases~~;~~ . ~~additionally, i~~In the ~~is~~ particular case of a regression analysis between two standardized anomaly quantities, the Pearson correlation coefficient, R, represents not only a measure of the linear dependency of the two random ~~quantities~~ variables but also the slope of the linear relationship and a proxy of the difference and biases of the two datasets. In this respect, R can be seen as a good synthetic descriptor of the relationship between two standardized z-score datasets. The statistical significance of the existence of a positive correlation can be evaluated by means of the Student's t-~~student~~ test (2 sided) by computing the R value corresponding to a significance level $p = 0.05$.

Analysis of the correlation among the datasets is interesting in the framework of the triple collocation (TC) technique and its basic hypotheses. In TC, a first key hypothesis is the existence of linearity between the 'true' status of the system and the three models; this is formally expressed as:

$$z_x = \alpha_x + \beta_x z_\Theta + \varepsilon_x \qquad (2)$$

where z_Θ is the unknown true dataset of soil moisture anomalies, α_x and βx are the systematic slope and bias parameters for the dataset x with respect to the truth, and ε_x is the additive zero-mean random noise.

It follows that the absence of a statistical significant linear relation between all three models openly violates this hypothesis.

Other key underling hypotheses of TC are the stationarity of both signals and errors, the independency between the errors and the signal (error orthogonality) and the independence between the errors of the three datasets (zero-cross correlation) (Gruber et al., 2016). Finally, operational limitations regard the minimum sample size of each dataset, which is commonly assumed equal to 100 values (Scipal et al., 2008; Dorigo et al., 2010), even if some other authors suggest much larger sample sizes for a lower relative uncertainty (Zwieback et al., 2012).

Under these assumptions, Stoffelen (1998) proposed a formulation to estimate each model error variance, $\sigma^2_{\varepsilon x}$, based on a combination of the covariance between the datasets. In this approach, known as the covariance notation (Gruber et al., 2016), the error variance values are computed without a common (arbitrary) reference dataset as:

$$\begin{aligned} \sigma^2_{\varepsilon_1} &= \sigma^2_1 - \frac{\sigma_{12}\sigma_{13}}{\sigma_{23}} \\ \sigma^2_{\varepsilon_2} &= \sigma^2_2 - \frac{\sigma_{21}\sigma_{23}}{\sigma_{13}} \\ \sigma^2_{\varepsilon_3} &= \sigma^2_3 - \frac{\sigma_{31}\sigma_{32}}{\sigma_{12}} \end{aligned} \quad (3)$$

where, for the sake of simplicity, LIS, LST and CCI w~~h~~ere renamed 1, 2, 3, respectively. The first term on the right side of Eqs. (3) represents the single model data variance, whereas the second term represents the so-called sensitivity of the model to variations in the true status, which is a function of the covariance terms between the three models. The advantage of this formulation is to directly estimate the unscaled error variances, which can (eventually) be scaled to a common data space, if needed.

In the case of the application of the covariance notation to standardized quantities (with zero mean and unitary standard deviation), the error variance values computed through Eqs. (3) are expressed as dimensionless multiples of standard deviation, and a transformation to a common data space is not needed.

Different performance metrics can be derived from the covariance notation, including relative error variance metrics such as the fractional root-mean-squared-error (fRMSE, Draper et al., 2013) and the correlation coefficient of each model with the underlying true signal (McColl et al., 2014).

However, these metrics can be derived from each other by means of simple relationships (see Gruber et al., 2016) and they are analogous to the absolute error variance values in the case of z-scores ~~values~~ that have known unitary dataset variance.

3. Data and Materials

3.1 Lisflood model soil moisture

Root zone soil moisture dynamics are simulated by means of the Lisflood model (de Roo et al., 2000), a GIS-based distributed hydrological rainfall-runoff-routing model designed to reproduce the main hydrological processes that occur in large and trans-national European river catchments. The model simulates all the main hydrological processes occurring in the land-atmosphere system, including infiltration, actual evapotranspiration, soil water redistribution in three sub-layers (surface, root zone and sub-soil), surface runoff rooting to channel, and groundwater storage and transport (Burek et al., 2013).

Static maps used by the model are related to topography (i.e., digital elevation model, local drain direction, slope gradient, elevation range), land use (i.e., land use classes, forest fraction, fraction of urban area), soil (i.e., soil texture classes, soil depth), and channel geometry (i.e., channel gradient, Manning's roughness, bankfull channel depth, channel length, bottom width and side slope). Root zone depth is defined for each modelling cell on the basis of soil type and land use, where ~~T~~the soil-related hydraulic ~~quantities~~ properties are obtained from the ISRIC 1-km SoilGrids database (Hengl et al., 2014), whereas topography data are obtained from the Hydrosheds database (Lehner et al., 2008).

Daily meteorological forcing maps are derived from the European Centre for Medium-range Weather Forecasts (ECMWF) data as spatially resampled and harmonized by the JRC Monitoring Agricultural ResourceS (MARS) group. The dataset includes daily average air temperature, potential evapotranspiration (for soil, water and reference surfaces) and total rainfall at 0.25 degree spatial resolution, which were resampled on the model grid using the nearest neighbors algorithm.

The model run used in this study includes daily maps at 0.1 degree resolution between 1989 and 2015; the grid domain of this dataset is used as reference for the other two, whereas the baseline for the anomalies computation is defined by the period 2001-2015 in order to match the LST data availability. Monthly data to be used in Eq. (1) are computed as a simple average of all the data available for each month, given that no gaps can be found in this dataset due to its continuous nature as hydrological model. However, some areas where masked out due to the minimum or null temporal dynamic of soil moisture, such are Greenland and the Sahara desert.

3.2 Land Surface Temperature dataset

The use of the land surface temperature (LST) anomalies as a proxy of soil moisture anomalies is based on the well-known role of LST in the surface energy budget as a control factor for the partitioning between latent and sensible heat fluxes. In recent years, the existence of a connection between soil moisture and LST has been analyzed, mainly through the thermal inertia and the triangle methods (e.g., Carlson 2007; Verstraeten et al., 2006), as well as a direct proxy (see e.g., Park et al., 2014; Srivastava et al., 2016). In a study over the pan-European domain, Cammalleri and Vogt (2016) have demonstrated the good agreement between monthly LST and LIS-based root zone soil moisture z-score values during summer time, where LST outperforms other LST-based indicators such as the day-night difference and the surface-air gradient.

Following these findings, this study adopts the dataset collected by the Moderate-Resolution Imaging Spectroradiometer (MODIS) sensor on board of the Terra satellite (http://terra.nasa.gov/about/terra-instruments/modis) as a source of monthly-scale long records of LST maps. In particular, the MOD11C3 Monthly CMG (Climate Modelling Grid) LST product is used in this study, which is constituted by monthly composited and averaged temperature and emissivity maps at a spatial resolution of 0.05 degrees over a regular latitude/longitude grid; data for the period 2001–2015 are used, as the only fully completed years at the time of the analysis.

This monthly composite product is obtained as an average of the clear-sky data in the MOD11C1 products on the calendar days of the specific month, which are derived after a re-projecting and a re-sampling of the MOD11B1 product. Details on the algorithms used to obtain the daily MOD11B1 maps can be found in Wan et al. (2002); in summary, a double screening procedure is applied, based on: i) the difference between the two independent LST estimates of the day/night algorithm (Wan and Li, 1997) and the generalized split-window algorithm (Wan and Dozier, 1996), and ii) the histogram of the difference between daytime and nighttime LSTs.

LST monthly maps were spatially co-registered to the Lisflood 0.1 degree regular latitude/longitude grid by means of a simple average of the values within each cell, and anomaly maps were computed according to Eq. (1) by using only the data for which LST > 1 °C; this threshold value (commonly used in snowmelt and snow/rainfall discrimination procedures; WMO, 1986) allows removing the data that are likely affected by snow/frost from the analysis.

3.3 Microwave combined dataset

The ESA Climate Change Initiative (CCI) aims at developing a multi-satellite soil moisture dataset by combining data collected in both past and present by passive and active microwave instruments (~~Dorigo et al., 2016~~Liu et al., 2012; Wagner et al., 2012). The current version of the dataset (v03.2) combines data from nine different sensors (SMMR, ERS-1/2, TMI, SSM/I, AMSR-E, ASCAT, WindSat, AMSR2 and SMOS) between 1978 and 2015.

Satellite-based microwave estimates of soil moisture are usually related to the first few centimeters of soil column (i.e., skin layer), which is quite closely related to the soil moisture content in the root zone (Paulik et al., 2014), except for very dry conditions in sandy soils. Additionally, numerous validations against land surface models have highlighted good performance across the globe, with notable exceptions over densely vegetated areas (e.g., Loew et al., 2013).

The algorithm adopted to merge the different data sources is the one developed by Liu et al. (2012), which is a three-step procedure that: i) merges the original passive microwave products, ii) merges the original active microwave products, and iii) blends the two merged products into a single final dataset. The merging procedure of passive datasets includes pixel-scale separation between seasonality and anomalies, rescaling of the data based on the piece-wise cumulative distribution function (CDF) and merging of the dataset using a common reference seasonality. For the active microwave instruments, the CDFs are directly used to rescale the data under the assumption that active datasets have an identical dynamic range, this mainly due to the limited overlap between datasets. The final blending of the two merged datasets is obtained by adopting a common resolution of approximately 25 km and daily frequency, as well as by using the GLDAS-1-Noah model (ftp://agdisc.gsfc.nasa.gov/data/s4pa/) as a reference dataset for the CDF matching.

In this study, the daily blended dataset is spatially resampled to a 0.1 degree regular latitude/longitude grid (the same used in Lisflood simulations) by means of the nearest neighbor algorithm, and successively aggregated to monthly time scale by simply averaging the data (only if at least 8 daily values were available in the specific month). Monthly average maps were converted into z-score maps by using the baseline period 2001-2015 (the timeframe available for the LST dataset). Monthly aggregated z-score values of skin soil moisture are analyzed, jointly with the other two datasets, under the assumption that time-aggregation and normalization procedures minimize some of the discrepancies that are likely present between skin and root zone daily time-series.

4. Results and Discussion

Considering the assumption of linearity between each one of the ~~models~~ datasets and the unknown true status of the system in TC, a preliminary analysis on the linear correlation between the three anomaly ~~models~~ products has been performed in order to detect the macro-areas where the TC procedure can be applied without violating this basic hypothesis. The correlation analysis was performed by using only the monthly anomalies ~~data~~ that were available for all three ~~models~~datasets, with at least a sample size of 100 values (max sample size = 12 months × 15 years = 180), and by defining a minimum correlation threshold ($R_{0.05}$) that ensures a statistical significance of the linear relationship on the basis of the Student's t-~~student~~ test (at $p = 0.05$).

The map in Fig. 1 reports in grey the areas where all three ~~models~~ datasets are significantly linearly correlated according to the described criteria, representing the areas where the first basic hypothesis of the TC is not clearly violated. It is worth to point out that some areas are excluded from the analysis by the lack of data in LIS (low temporal variability, as over Greenland and the Sahara desert), LST (due to the minimum temperature threshold or low temporal variability) or CCI (densely vegetated areas, such as the Amazon forest and the Congo basin). These results suggest to focus the successive detailed analysis on five macro-regions (demarked by the boxes in Fig. 1) that have consistent positive correlation values for all the three ~~models~~datasets; these areas are named, from now on, as: 1) NA (including the contiguous U.S. and Mexico), 2) EU (Southern and Central Europe), 3) SA (Southern countries of the African continent and Madagascar), 4) IN (Indian subcontinent), and 5) AU (Australia)*.

The correlation coefficient maps over those regions, obtained by inter-comparing the three ~~models~~datasets, are reported in Figs. 2 to 4, where the cells in red and yellow are the ones with negative or not-significant correlation, respectively, whereas the blue scale represents the cells with increasing significant linear correlation (from light to dark tones). The comparison between LIS and LST (Fig. 2) shows an overall good agreement between the two datasets, with only minor areas characterized by negative/not-significant correlation values; notably, low correlation can be observed over the Great Lakes and Rocky mountain areas in the U.S., over the Alps in Europe, North Angola and Western Himalaya. Similar results can be observed in Fig. 3, where LIS and CCI datasets are compared; this comparison shows an increasing number of negative values in the Western U.S., the Alps, and Southern Turkey, but overall high correlation values across most of the five regions. Finally, the comparison between LST and CCI reported in Fig. 4 shows an increase of areas with low/not-significant correlation in the Eastern and Western U.S. and both North- and South-Eastern Europe and the Alps, whereas a high correlation can be observed all over the other regions.

* Consider the countries and boundaries reported here only as indicative of the interested areas, and they may not in any circumstances be regarded as stating an official position of the European Commission.

On average, the data in Table 1 summarize the results obtained for all the regions together, as well as for each region independently, showing how CCI and LST are the two datasets best correlated to each other overall, even if this result is mainly driven by the results over AU, SA and IN macro-areas. The data of the LIS model are similarly correlated to the ones of LST and CCI, with a more uniform distribution of the results across the various sub-regions. Another outcome of this analysis is that the area with the lowest average correlation between the three ~~models~~ datasets is the EU, probably due to the high heterogeneity of this region at the 0.1 degree spatial scale.

Some of the discrepancies observed in Figs. 2 to 4 can be explained by the differences in both horizontal and vertical resolution of the three raw datasets. LIS is characterized by an higher spatial resolution (5-km) compared to CCI (25-km) and a vertical resolution that encompasses the full root zone against the skin soil moisture of the latter; LST has the same spatial resolution of LIS but a vertical resolution that varies as function of the vegetation coverage between skin (for bare soil) to root zone (for full vegetation coverage). The impact of such differences is partially reflected in the observed results, with CCI-LST better related over shallow soil in homogeneous areas, and LIS-LST better in agreement over sparse agricultural areas in Europe. Overall, it seems that the adopted expedients (i.e., monthly average, standardization) successfully minimized these issues, given that the results in Table 1 shows a substantial and similar agreement of the three datasets in the main areas.

~~Overall~~Additionally, the obtained results seem to suggest that it is reliable to adopt ~~use of~~ LST anomalies as proxy of soil moisture anomalies ~~seems based on a reliable assumption~~, since there is a clear consistency of LST anomalies with the other two datasets. Similar results were obtained by Fang et al. (2016) over the ~~C~~continental United States, where the outputs of the thermal-based ALEXI (Atmosphere Land EXchange Inverse) model compare~~s~~ well with soil moisture anomalies from CCI and Noah land-surface model. This consideration allows applying the TC analysis to the LST dataset as well, whereas most of the studies in the literature focus on land modelled and microwave soil moisture datasets (i.e., Dorigo et al., 2010; Gruber et al., 2016; Su et al., 2014) with only few notable exceptions including thermal data (e.g., Hain et al., 2011; Yilmaz et al., 2012).

The outputs of the correlation analysis were used to detect the cells suitable for the TC technique; since a key hypothesis of the technique is the existence of a linear relation between each model and the (unknown) truth, a necessary condition (even if not sufficient) is the existence of linear relationships among the three ~~models~~datasets. As outcome of the correlation analysis, around 10% of the five macro-areas were removed from the TC analysis due to the absence of this basic condition.

The maps in Figs. 5 to 7 show the main outcome of the TC analysis, which is the spatial distribution of the error variance (dimensionless, ~~as~~showing the multiple of model standard deviation) for each model, as detailed by Eqs. (3). The blank areas in those maps correspond to the cells where no significant linear correlation was observed between all three ~~models~~datasets. The results for LIS (Fig. 5) show how the highest errors are observed over the Western U.S., Northern Cape in South Africa and Western/Southern Australia, whereas low errors are observed over the Eastern U.S. On the opposite, the LST dataset displays the highest errors over the latter area (Fig. 6), whereas the lowest errors are observed over Queensland in Australia, Eastern Cape in South Africa and Lesotho. The maps in Fig. 7 show that the CCI dataset has consistent patterns of low error variance values over most of Australia, Western India and Central U.S.

Overall, on the one hand, it seems evident how CCI tends to outperform the other two methods over dry areas such as Australia and South Africa, but on the other hand, a region like the U.S. is almost equally subdivided among the three ~~models~~datasets, where LIS performs better in the East, LST in the West and CCI in the center. Differences among ~~models~~ products can be partially explained by the differences in the soil layer monitored by each dataset, i.e., microwave system capturing skin soil moisture whereas Lisflood models the full root zone; indeed, even if the use of monthly anomalies allows minimizing some of the discrepancies, skin soil moisture remains more reliable for dry/bare areas (Das et al., 2015). Even if these considerations partially explain the agreement/disagreement ~~behavior~~ of the three datasets, it is not straightforward to pinpoint in detail~~s~~ climate and/or vegetation derived patterns in the spatial distribution of the TC outputs.

These findings are summarized in the data reported in Table 2, where the average error variance for each model and macro-area is reported aside its spatial standard deviation. The data in Table 2 confirm that CCI has an overall better performance (lower errors) than LIS and LST, which perform quite closely, mainly thanks to the very low error variance observed over Australia and, to a minor extend, Southern Africa. The LIS model shows to perform better over NA and EU regions, likely due to the better meteorological forcing datasets available over those regions compared to the other macro-areas (due to denser ground networks). The LST dataset seems to perform moderately well over all five macro-regions, with the only notable exception of EU; however, it rarely outperforms the other two datasets, constituting a "second-best" option in most of the cases. It is also worth to point out that the CCI dataset is often masked-out over those regions where the error of microwave techniques are likely high, whereas the data of the other two datasets are mostly produced globally; hence, a possible explanation of the better performance of CCI compared to LIS and LST may be linked to this preliminary screening of the data.

The outcome that LIS slightly outperforms the other two datasets over NA is in agreement with the results reported by Hain et al. (2011), where the Noah land-surface model slightly outperforms (on average) the microwave and thermal datasets over Contiguous U.S. However, it should be pointed out how the spatial distribution of the error estimates for LIS differs from the ones reported for Noah, likely due to the differences in both meteorological forcing and modelling approaches. Some qualitative analogies can ~~be~~ also be observed with the results reported in ~~Similarly,~~ Pierdicca et al. (2015), which show~~s~~ smaller average errors at daily time scale over Europe for the ERA-LAND modelled datasets compared to two microwave-based datasets, even if both the temporal scale and the adopted methodology of the latter differ from the ones used in ~~similarly to the results obtained in this~~ our study. ~~Both t~~This~~ese~~ previous study ~~results~~ seems to ~~suggest~~ confirm that land modelling approaches are more reliable, on average, over these regions, likely due to the reliability of meteorological forcing and model parameterizations, even if there can be significant differences among the performances of different land-surface models.

Over AU sub-region, the spatial distribution of the errors in CCI are quite in agreement with the results reported in Su et al. (2014) for two microwave datasets, with larger errors along the South-East Australian coast. This result supports the assumption that microwave data are more reliable over dry bare soil areas, which is further highlighted by the results obtained in SA and IN sub-regions. The subdivision of the NA domain in three main regions is similar to the one observed by Gruber et al. (2016) in comparing ASCAT and AMSR-E microwave datasets, suggesting key differences in the soil moisture behavior over these three sub-regions. Overall, the spatial patterns of microwave and land model errors show similarities with the ones observed by Dorigo et al. (2010), even if no thermal data were included in their analysis.

The error variance values can also be interpreted as the correlation coefficient of each dataset with the underlying true signal, following the definition of McColl et al. (2014). In fact, for the special case of anomalies with unitary variance ($\sigma^2_x = 1$), the TC-derived R_x of each dataset is simply equal to $\sqrt{1-\sigma^2_{\varepsilon_x}}$, which ranges on average over all five regions (not shown) between 0.91 (for CCI in AU) to 0.66 (for LST over EU); these values show a good capability of the ~~models~~ datasets to capture, on average, temporal variations in soil moisture anomalies.

In order to provide a simple synthetic representation of the likely best model for each area, the map in Fig. 8 depicts for each cell the dataset with the lowest error variance by associating different colors to the three ~~models~~ datasets (red for LIS, blue for LST and green for CCI). Even if this approach is rather simplistic, as it cannot account for two ~~models~~ products performing really close over some areas, the major relevant features, like the predominance of the CCI model over Australia, are made evident by these maps.

The maps in Fig. 8 confirm CCI as the dataset with the lowest error variance values over most of AU, SA and IN, whereas the three ~~models~~ datasets almost equally split the other two macro-areas; this is even more evident in the data reported in Table 3, where the percentage of sub-areas where each model is the best is reported. These data confirm the good performance of CCI over AU, SA and IN macro-regions, whereas the NA territory is almost equally divided among the three datasets and LIS outperforms both LST and CCI over 50% of EU domain. In the latter, the areas where the LIS dataset outperforms the other two datasets partially resemble the results obtained by Pierdicca et al. (2011) for the ERA-LAND model; however, the present study includes also remote sensing thermal data and not only microwave-derived datasets. Overall, the CCI dataset outperforms the other two datasets in about 50% of cells, with the remaining almost equally split between LIS and LST.

Finally, the spatial distribution of the weighting factor of each dataset, computed according to the least square theory (Yilmaz et al., 2012), is represented in Figs. 9 to 11. The color scale of the figures was designed to represent in a neutral color the cells that have a weighing factor close to the one for a simple-average (1/3), in green scale the weights greater than a simple-average (larger contribute) and in orange the weights lower than the simple-average (smaller contribute). The visual intercomparison of the three maps further emphasizes the good performance of the CCI product over AU and SA, the best performance of LIS over the Eastern US and EU, and the good results obtained for LST in Western US and Northern AU. It is worth noticing that the use of a weighted average based on the TC error analysis does not seem to bring advantages over large areas of central US, EU and Eastern IN where the weighting factors are close to the ones for a simple arithmetic average.

The behavior of the weighting factors over the five macro-areas can be synthetized by the frequency diagram in Fig. 12. This plot shows the high fraction of weighting factors > 0.4 for the CCI dataset, representing a predominant contribution on the ensemble mean of this product over the others, whereas LST has a peak of frequency center around 1/3 (arithmetic average) and LIS has a hint of bi-modal distribution. These data, together with the map in Fig. 8, confirm the fact that CCI outperforms the other two datasets in 50% of the domain, whereas LST is often the second-best option behind either CCI or LIS.

5. Summary and Conclusions

Three datasets have been compared as proxy of the unknown true status of soil moisture anomalies in the context of the global drought monitoring system under development by the JRC of the European Commission. Key assumption of the study is the inability of a single dataset to accurately capture the soil moisture dynamic over the large range of variability of conditions that can be observed at continental to global scale.

The inter-comparison between the three datasets, namely the outputs of the Lisflood hydrological model (LIS), the MODIS-based land surface temperature (LST) and the combined active/passive satellite microwave (CCI) data, confirms some inconsistencies between the three datasets over ~~some~~ certain areas, as well as the difficulties in comparing the three datasets over peculiar areas (e.g., Sahara desert, Amazon rainforest) due to the lack of coverage from one or more datasets. Generally, the three datasets seem comparable over most of the globe, thanks to the use of time-aggregation and standardization procedures that remove temporal inconsistencies and biases among the series. Focusing the analysis only on the areas where the three ~~models~~ datasets are substantially in agreement (following a linear regression analysis), five macro-regions were detected as suitable for further investigations according to the Triple Collocation (TC) technique. This analysis allows quantifying the likely random error associated with each model (with regard to the true status) even in absence of an observation of the "truth", under the hypothesis that certain criteria are met.

The main outcome of the TC analysis further confirms the need of a multi-source approach for a reliable assessment of soil moisture anomalies over those five regions, given that no model outperforms the others (in terms of expected error variance) for the entire study domain. Emblematic are the results over North America, where each model outperforms the others in one sub-region, like the LIS approach in Eastern U.S., LST in the Southern-Western domain and CCI in Central U.S. Even if no clear insight on the general patterns of the errors can be provided as outcome of the study, ~~O~~overall, the obtained results seem suggesting that remote sensing datasets ~~seem to~~ perform better over dry areas and sparsely monitored areas (e.g., Australia and Southern Africa), whereas the LIS dataset seems more reliable over NA and EU where dense networks of meteorological ground stations are deployed.

It has been highlighted how some differences among ~~models~~ the datasets can also be related to the depth of the soil layer monitored by each dataset, i.e., the microwave system capturing skin soil moisture whereas Lisflood models the full root zone; indeed, even if the use of monthly anomalies allows minimizing some of the discrepancies and biases, our results confirm that skin soil moisture remains more reliable for ~~dry/bare~~ areas where the effects of vegetation coverage is minimal (Das et al., 2015), whereas hydrological models are more suited for agricultural and densely vegetated regions.

However, the three datasets seems to be overall comparable in terms of average performances, supporting the success of the adopted homogenization procedures. Some analogies between the obtained results and the ones already available in the literature have been found, but the inclusion of thermal data into the analysis enlarges the understanding of the mutual relationship between the different datasets.

The results of this study represent a robust starting point for the development of a global drought monitoring system based on such anomaly datasets, which can exploit the main findings of the TC analysis in order to develop a suitable ensemble product over the investigated regions. The error characterization derived from TC was used to estimate the weighing factors of an ensemble mean procedure, based on the least squares framework reported in Yilmaz et al. (2012). Currently, an operational implementation of such ensemble product is foreseen for the GDO system as soon as the CCI product becomes available in near-real time.

Further analyses are required to be able to extend the test to the areas currently not included in this study, especially the ones where the three datasets are available but provide inconsistent or contrasting results. In this context, the analysis of further global datasets may help in unveil the reasons behind such discrepancies.

References

Anderson, M.C., Norman, J.M., Mecikalski, J.R., Otkin, J.P., Kustas, W.P., 2007. A climatological study of evapotranspiration and moisture stress across the continental U.S. based on thermal remote sensing: II. Surface moisture climatology. J. Geophys. Res. 112, D11112, doi:10.1029/2006JD007507.

Burek, P., van der Knijff, J.M., de Roo, A., 2013. LISFLOOD: Distributed Water Balance and Flood Simulation Model. JRC Scientific and Technical Reports, EUR 26162 EN, 142 pp. doi:10.2788/24719.

Cammalleri, C., Vogt, J.V., 2016. On the role of Land Surface Temperature as proxy of soil moisture status for drought monitoring in Europe. Remote Sens. 7, 16849-16864.

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.

Campbell, G.S., Norman, J.M., 1998. An introduction to environmental biophysics, Springer-Verlag., New York (NY), USA. doi: 10.1007/978-1-4612-1626-1.

Carlson, T., 2007. An overview of the “Triangle Method” for estimating surface evapotranspiration and soil moisture from satellite imagery. Sensors 7(8), 1612-1629.

Crow, W.T., Kumar, S.V., Bolten, J.D., 2012. On the utility of land surface models for agricultural drought monitoring. Hydrol. Earth Syst. Sci. 16, 3451-3460.

Dai, A., 2011. Drought under global warming: A review. Wiley Interdiscip. Rev. Clim. Change 2, 45-65.

Das, K., Paul, P.K., Dobesova, Z., 2015. Present status of soil moisture estimation by microwave remote sensing. Cogent Geoscience 1, 1084669.

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.

Dirmeyer, P.A., Gao, X., Zhao, M., Guo, Z., Oki, T., Hanasaki, N., 2006. GSWP-2: multimodel analysis and implications for our perception of the land surface. Bull. Amer. Meteor. Soc. 87, 1381–1397.

Dorigo, W.A., Scipal, K., Parinussa, R.M., Liu, Y.Y., Wagner, W., de Jeu, R.A.M., Naeimi, V., 2010. Error characterisation of global active and passive microwave soil moisture datasets. Hydrol. Earth Syst. Sci. 14, 2605-2616.

~~Dorigo, W.A., Chung, D., Gruber, A., Hahn, S., Mistelbauer, T., Parinussa, R.M., Paulik, C., Reimer, C., van der Schalie, R., de Jeu, R.A.M., Wagner, W., 2016. Soil moisture [in "State of the Climate in 2015]. Bull. Amer. Meteor. Soc., 97(8), S31-S32.~~

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.

Fang, L., Hain, C.R., Zhan, X., Anderson, M.C., 2016. An inter-comparison of soil moisture data products from satellite remote sensing and a land surface model. Int. J. Appl. Earth Obs. Geoinf. 48, 37-50.

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. Geoinf. 45, 200-211.

Hain, C.R., Crow, W.T., Mecikalski, J.R., Anderson, M.C., Holmes, T., 2011. An intercomparison of available soil moisture estimates from thermal infrared and passive microwave remote sensing and land surface modeling. J. Geophys. Res. 116, D15107.

Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B.M., Ribeiro, E., et al., 2014. SoilGrids1km — Global Soil Information Based on Automated Mapping. PLoS ONE 9(8), e105992.

Houborg, R., Rodell, M., Li, B., Reichle, R., Zaitchik, B., 2012. Drought indicators based on model assimilated GRACE terrestrial water storage observations. Wat. Resour. Res. 48, W07525. doi:10.1029/2011WR011291.

Jackson, T.J., 2006. Estimation of Surface Soil Moisture Using Microwave Sensors. Encyclopedia of Hydrological Sciences, Part 5: Remote Sensing. doi: 10.1002/0470848944.hsa060.

Lehner, B., Verdin, K., Jarvis, A., 2008. New global hydrography derived from spaceborne elevation data, Eos 89(10), 93–94.

Liu, Y.Y., Dorigo, W.A., Parinussa, R.M., de Jeu, R.A.M., Wagner, W., McCabe, M.F., Evans, J.P., van Dijk, A.I.J.M., 2012. Trend-preserving blending of passive and active microwave soil moisture retrievals. Remote Sens. Environ. 123, 280-297.

Loew, A., Stacke, T., Dorigo, W., de Jeu, R., Hagemann, S., 2013. Potential and limitations of multidecadal satellite soil moisture observations for selected climate model evaluation studies. Hydrol. Earth Syst. Sci. 17, 3523-3542.

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. Let. 41, 6229-6236.

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, 1423-1429.

Mo, K.C., Long, L.N., Xia, Y., Yang, S.K., Schemm, J.E., Ek, M.B., 2010. Drought indices based on the Climate Forecast System Reanalysis and ensemble NLDAS. J. Hydrometeorol. 12, 185-210.

Park, J.-Y., Ahn, S.-R., Hwang, S.-J., Jang, C.-H., Park, G.-A., Kim, S.-J., 2014. Paddy Water Environ. 12(1), 77-88.

Paulik, C., Dorigo, W., Wagner, W., Kidd, R., 2014. Validation of the ASCAT Soil Water Index using in situ data from the International Soil Moisture Network. Int. J. Appl. Earth Obs. Geoinfo. 30, 1-8.

Pierdicca, N., Fascetti, F., Pulvirenti, L., Crapolicchio, R., Munõz-Sabater, J., 2015. Analysis of ASCAT, SMOS, in-situ and land model soil moisture as a regionalized variable over Europe and North Africa. Remote Sens. Environ. 170, 280-289.

Price, J.C., 1980. The potential of remotely sensed thermal infrared data to infer surface soil moisture and evaporation. Water Resour. Res. 16(4), 787-795.

Scipal, K., Holmes, T., de jeu, R., Naemi, V., Wagner, W., 2008. A possible solution for the problem of estimating the error structure of global soil moisture datasets. Geophys. Res. Lett. 35, L24404. doi:10.1029/2008GL035599.

Sheffield, J., Goteti, G., Wen, F., Wood, E.F., 2004. A simulated soil moisture based drought analysis for the United States. J. Geophys. Res. 109, D24108. doi:10.1029/2004JD005182.

Stoffelen, A., 1998. Toward the true near-surface wind speed: Error modelling and calibration using triple collocation. J. Geophys. Res. 103, 7755-7766.

Srivastava, P.K., Islam, T., Singh, S.K., Gupta, M., Petropoulos, G.P., Gupta, D.K., Wan Jaafar, W.Z., Prasad, R., 2016. Soil moisture deficit estimation through SMOS soil moisture and MODIS land surface temperature. In: Satellite Soil Moisture Retrieval: Techniques and Applications, P.K. Srivastava, G.P. Petropoulos, Y.H. Kerr (Eds.), Elsevier B.V.

Su, C.-H., Ryu, D., Crow, W.T., Western, A.W., 2014. Beyond triple collocation: Applications to soil moisture monitoring. J. Geophys. Res. Atmos. 119, 6419-6439.

Verstraeten, W.W., Veroustraete, F., van der Sande, C.J., Grootaers, I., Feyen, J., 2006. Soil moisture retrieval using thermal inertia, determined with visible and thermal spaceborne data, validated for European forests. Remote Sens. Environ. 101(3), 299-314.

Wagner, W., Dorigo, W., de Jeu, R., Fernandez, D., Benveniste, J., Haas, E., Ertl, M., 2012. Fusion of active and passive microwave observations to create an essential Climate Variable data record on soil moisture. ISPRS Annal of the Photogrammetry, Remote Sensing and Spatial Information Sciences, volume I-7. XXII ISPRS Congress, 25 August – 01 September 2012, Melbourne, Australia.

Wan, Z., Zhang, Y., Zhang, Q., Li, Z.-L., 2002. Validation of the land-surface temperature products retrieved from Terra Moderate Resolution Imaging Spectroradiometer data. Remote Sens. Environ. 83, 163-180.

Wan, Z., Li, Z.-L., 1997. A physics-based algorithm for retrieving land-surface emissivity and temperature from EOS/MODIS data. IEEE Trans. Geosci. Remote Sens. 35, 980-996.

Wan, Z., Dozier, J., 1996. A generalized split-window algorithm for retrieving land surface temperature from space. IEEE Trans. Geosci. Remote Sens. 34, 892-905.

Wilhite, D.A., 2000. Drought as a natural hazard: Concepts and definitions. N: Disasters series. Routledge Publishers, UK, 213-230.

World Meteorological Organization, 1986. Intercomparison of models of snowmelt runoff. Operational Hydrological Report, 23.

Yilmaz, M.T., Crow, W.T., Anderson, M.C., Hain, C.R., 2012. An objective methodology for merging satellite- and model-based soil moisture products. Water Resour. Res. 48(11), W11502. doi:10.1029/2011WR011682, 2012.

Yilmaz, M.T., Crow, W.T., 2014. Evaluation of assumptions in soil moisture triple collocation analysis. J. Hydrometeorol. 15, 1293-1302.

Zwieback, S., Scipal, K., Dorigo, W., Wagner, W., 2012. Structural and statistical properties of the collocation technique for error characterization. Nonlin. Processes Geophys. 19, 69-80.

Tables

Table 1. Summary of the Pearson correlation coefficient values (average ± standard deviation) observed for all the regions.

Comparison	ALL	NA	EU	SA	IN	AU
LIS vs. LST	0.44 ± 0.09	0.41 ± 0.08	0.39 ± 0.07	0.48 ± 0.09	0.44 ± 0.07	0.50 ± 0.10
LIS vs. CCI	049 ± 0.10	0.47 ± 0.09	0.42 ± 0.08	0.48 ± 0.10	0.48 ± 0.08	0.58 ± 0.11
CCI vs. LST	0.56 ± 0.13	0.49 ± 0.14	0.37 ± 0.09	0.63 ± 0.09	0.52 ± 0.10	0.68 ± 0.07

Table 2. Summary of the TC error variance analysis, reporting the spatial average (± standard deviation) values observed over each macro-region.

Model	ALL	NA	EU	SA	IN	AU
LIS	0.48 ± 0.13	0.42 ± 0.14	0.44 ± 0.12	0.54 ± 0.11	0.49 ± 0.10	0.54 ± 0.14
LST	0.44 ± 0.13	0.46 ± 0.15	0.56 ± 0.10	0.37 ± 0.10	0.48 ± 0.09	0.38 ± 0.11
CCI	0.36 ± 0.18	0.46 ± 0.16	0.54 ± 0.12	0.30 ± 0.14	0.38 ± 0.16	0.17 ± 0.10

Table 3. Fraction of each macro-area (as percentage) where one model outperforms the other two.

Model	ALL	NA	EU	SA	IN	AU
LIS	25.5	39.2	50.0	10.6	28.2	4.3
LST	25.7	28.8	23.1	36.0	20.3	18.6
CCI	48.8	32.0	26.9	53.4	51.5	77.1

Figures

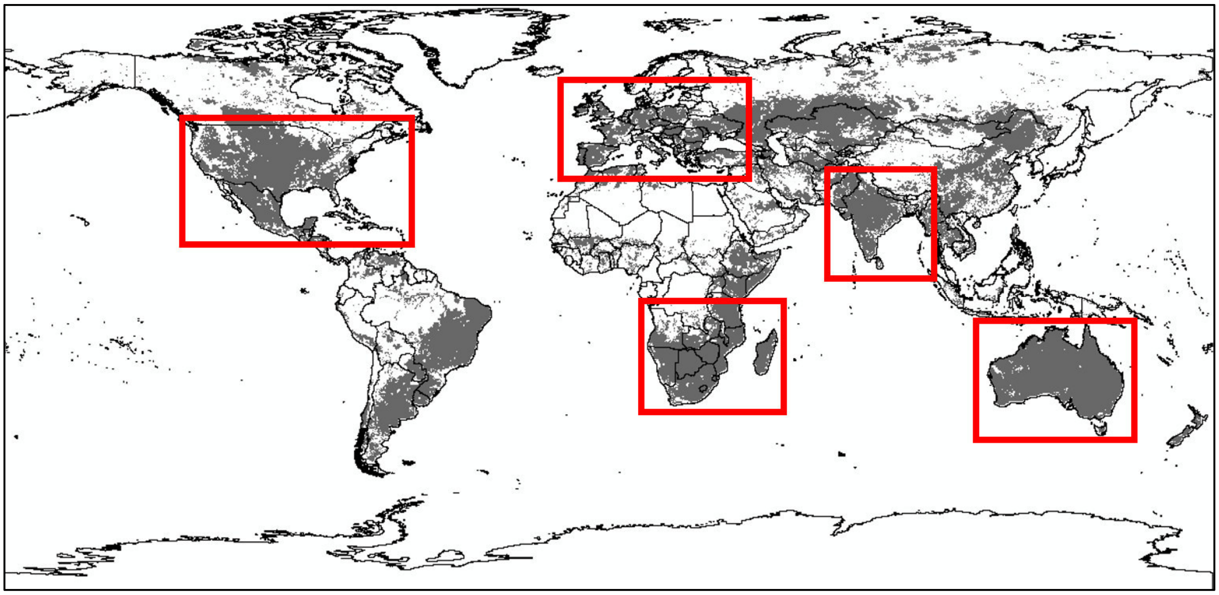

Fig. 1. Map of the areas where all the three models are positively significantly linearly correlated (cells in grey) according to the Student's t-student test at $p = 0.05$. The boxes delimitate the macro-regions selected for the successive analyses.

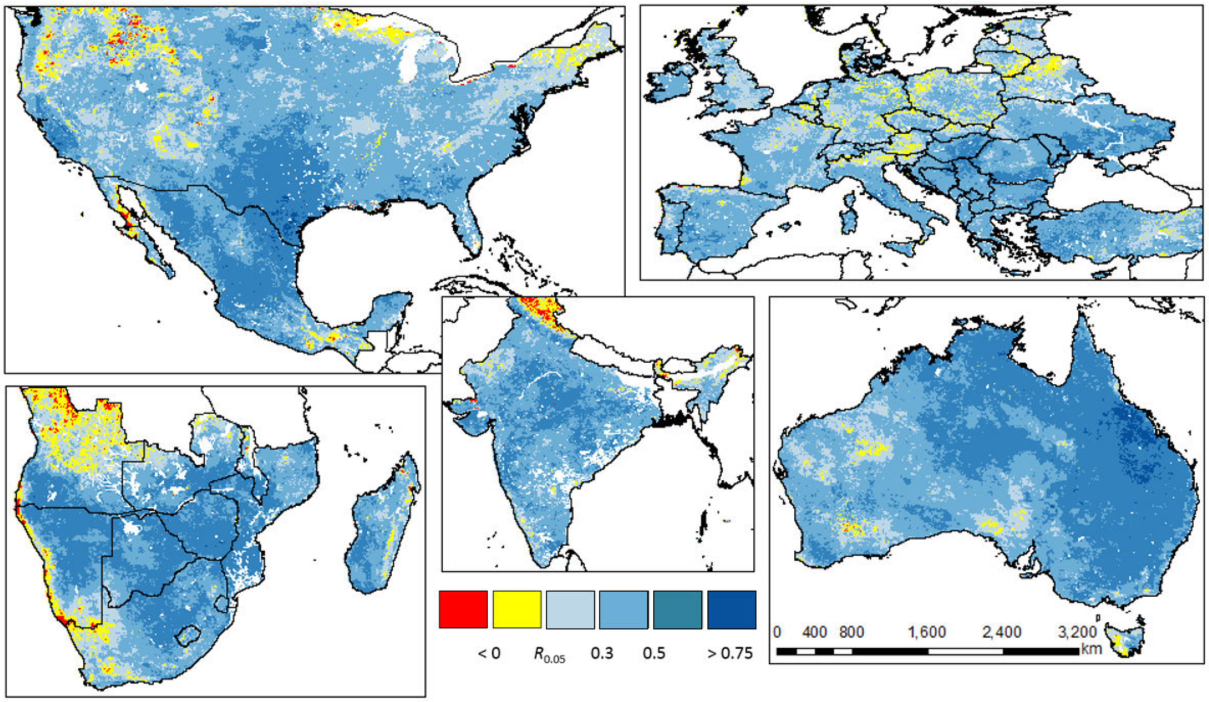

Fig. 2. Spatial distribution of the Pearson correlation coefficient (*R*) between Lisflood soil moisture anomalies (LIS) and land surface temperature anomalies (LST) over the five selected macro-regions. Values in red and yellow are negatively correlated or not significant at $p = 0.05$, respectively.

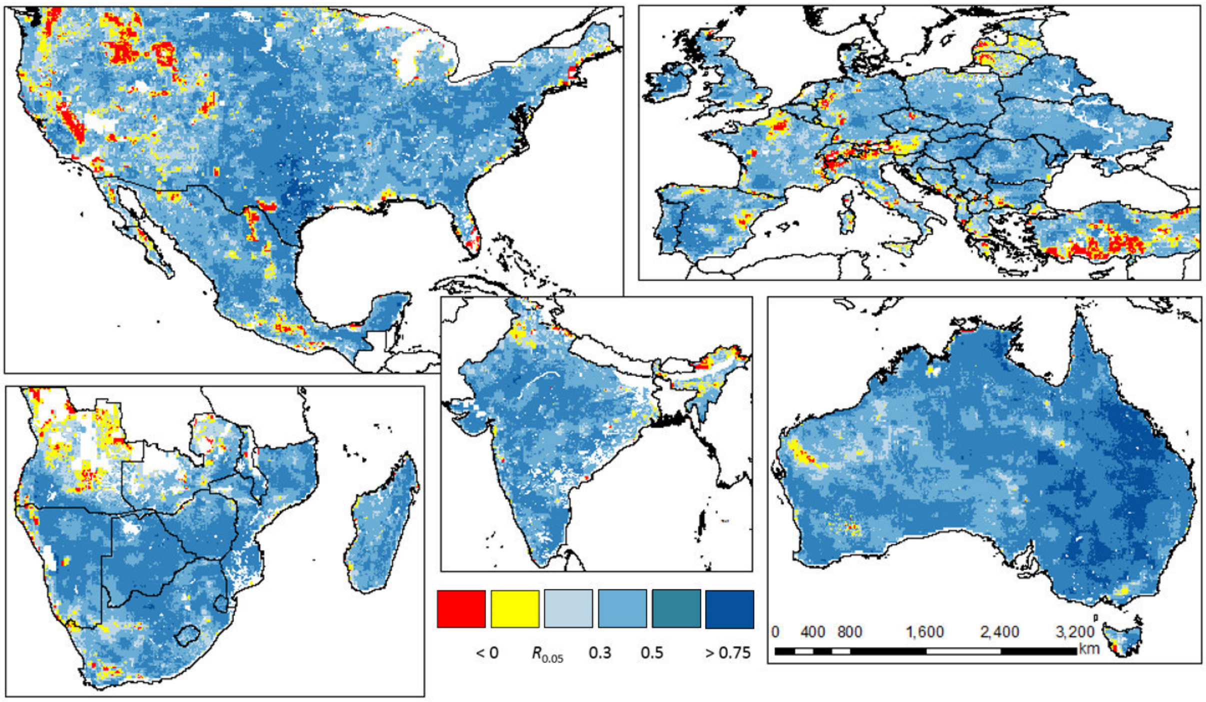

Fig. 3. Spatial distribution of the Pearson correlation coefficient (*R*) between Lisflood (LIS) and ESA Climate Change Initiative (CCI) soil moisture anomalies over the five selected macro-regions. Values in red and yellow are negatively correlated or not significant at $p = 0.05$, respectively.

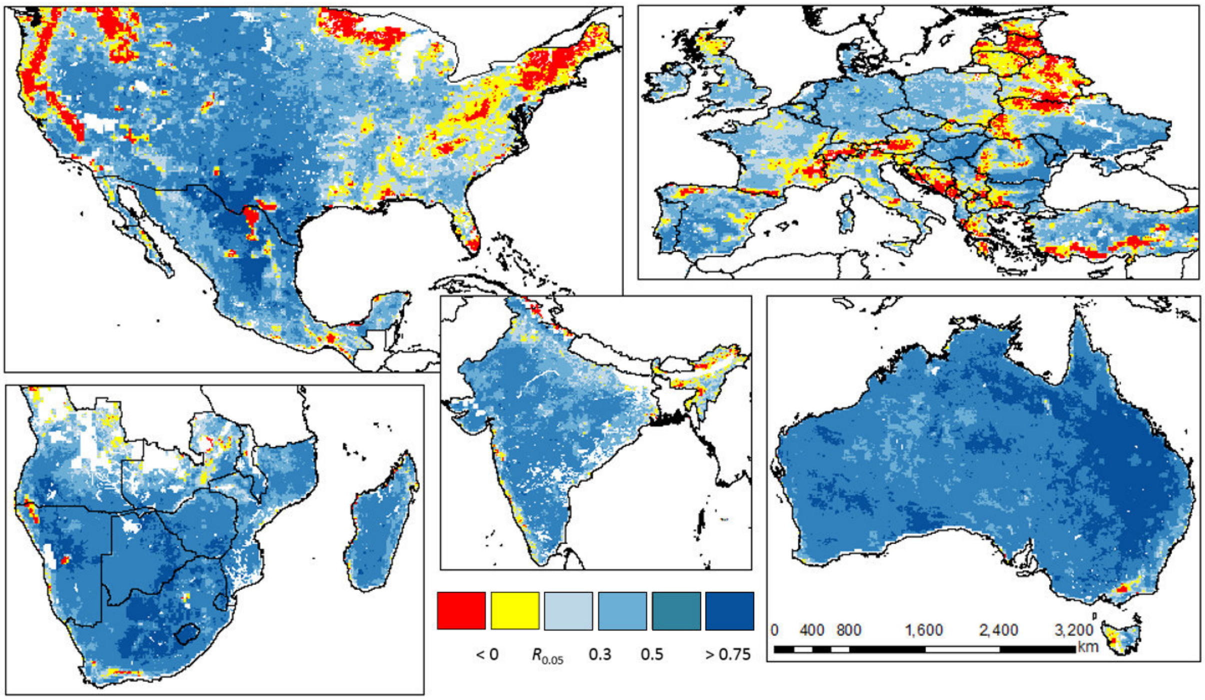

Fig. 4. Spatial distribution of the Pearson correlation coefficient (*R*) between ESA Climate Change Initiative soil moisture anomalies (CCI) and land surface temperature anomalies (LST) over the five selected macro-regions. Values in red and yellow are negatively correlated or not significant at $p = 0.05$, respectively.

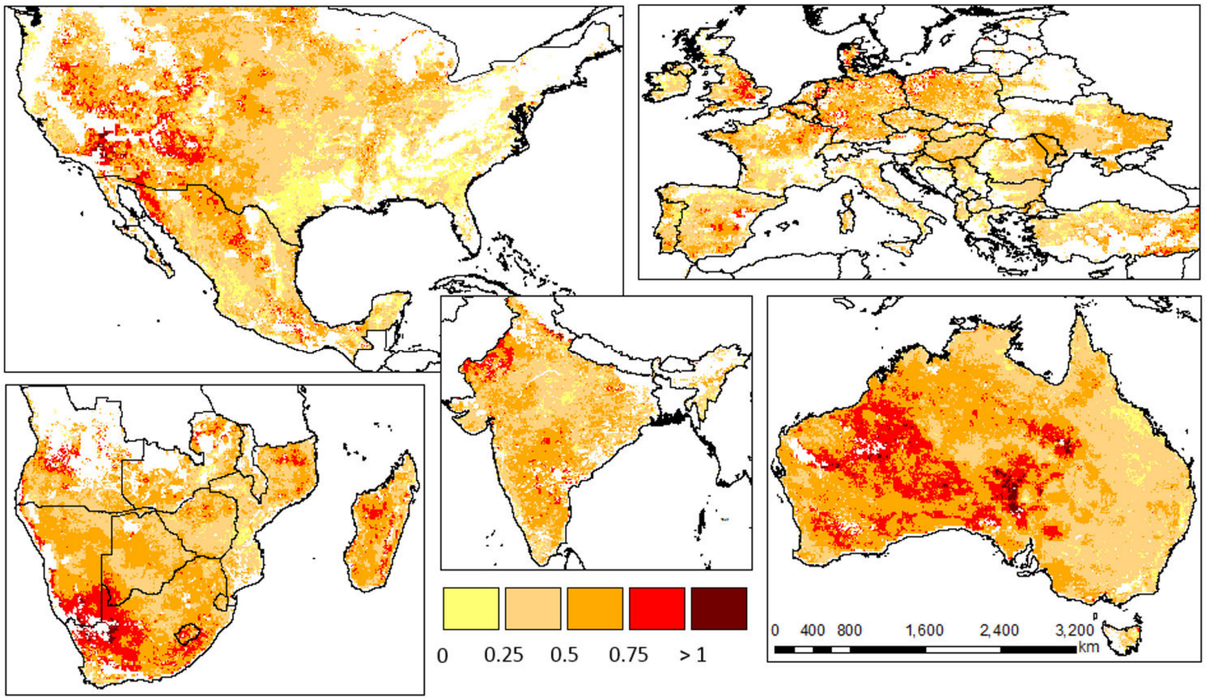

Fig. 5. Spatial distribution of the error variance for the Lisflood (LIS) dataset over the five selected macro-regions.

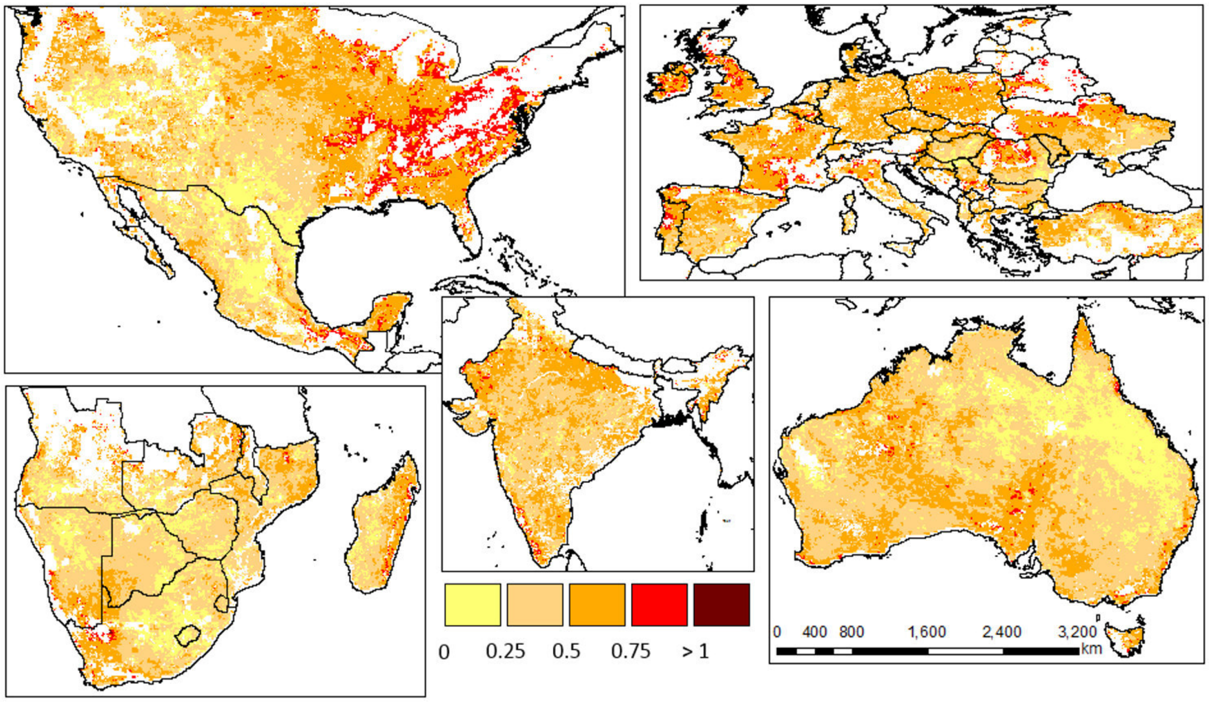

Fig. 6. Spatial distribution of the error variance for the land surface temperature (LST) dataset over the five selected macro-regions.

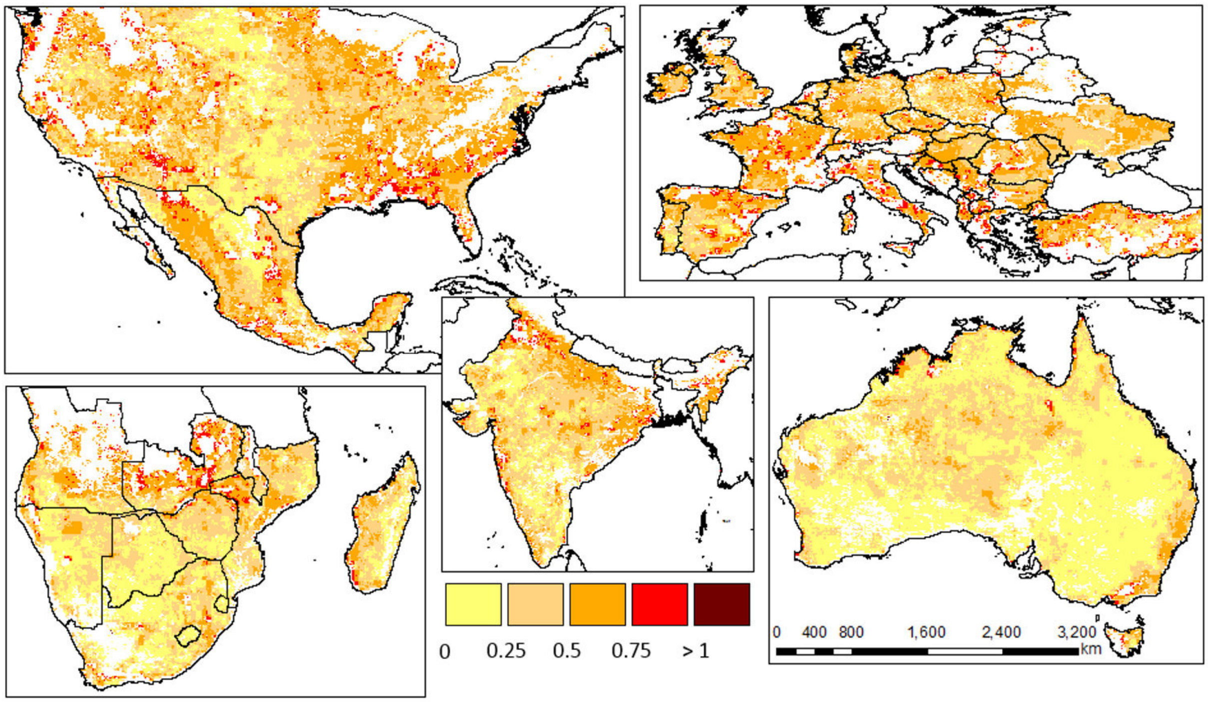

Fig. 7. Spatial distribution of the error variance for the ESA Climate Change Initiative (CCI) dataset over the five selected macro-regions.

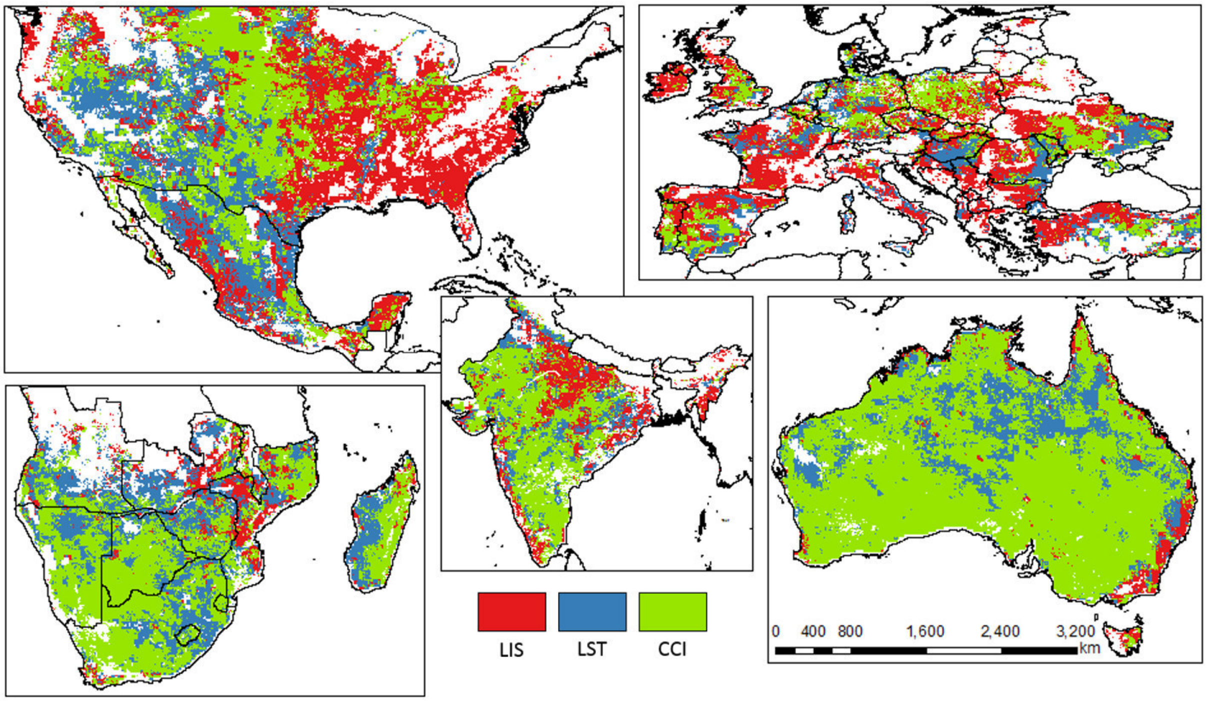

Fig. 8. Maps representing the best performing (lowest error variance) dataset for each cell according to the TC analysis.

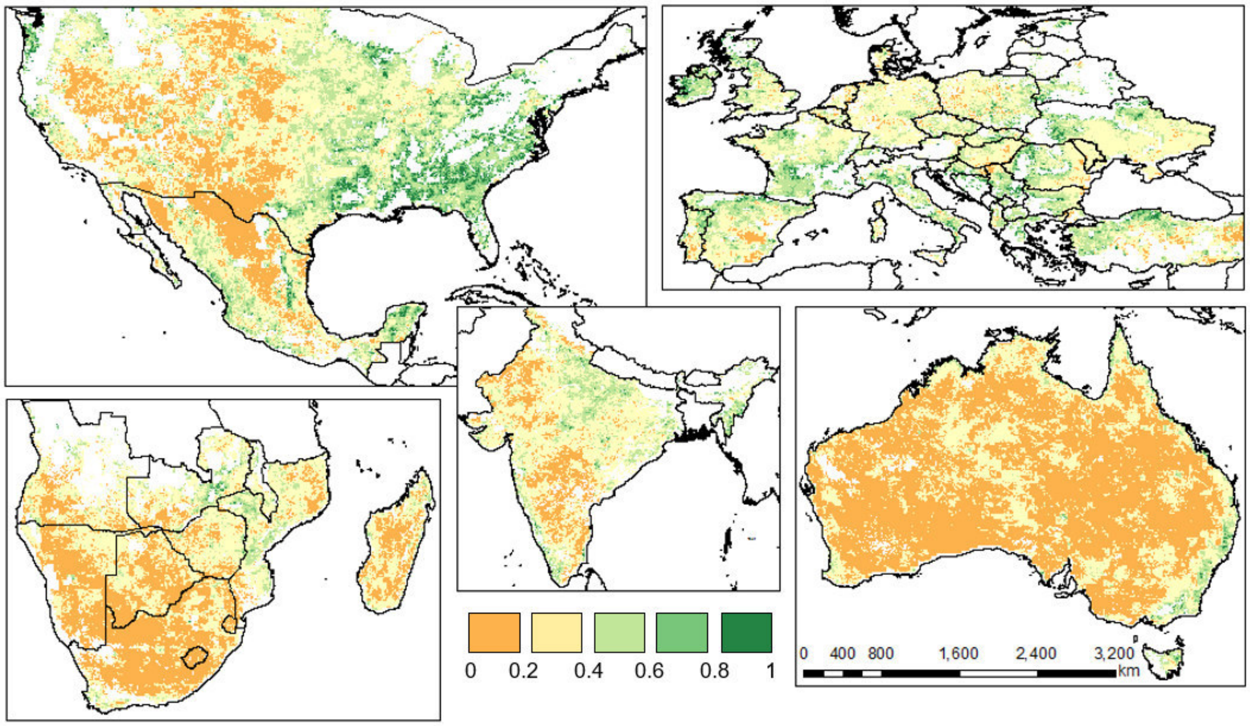

Fig. 9. Maps representing the ensemble mean weighting factor for the LIS dataset according to the error maps derived from the TC analysis.

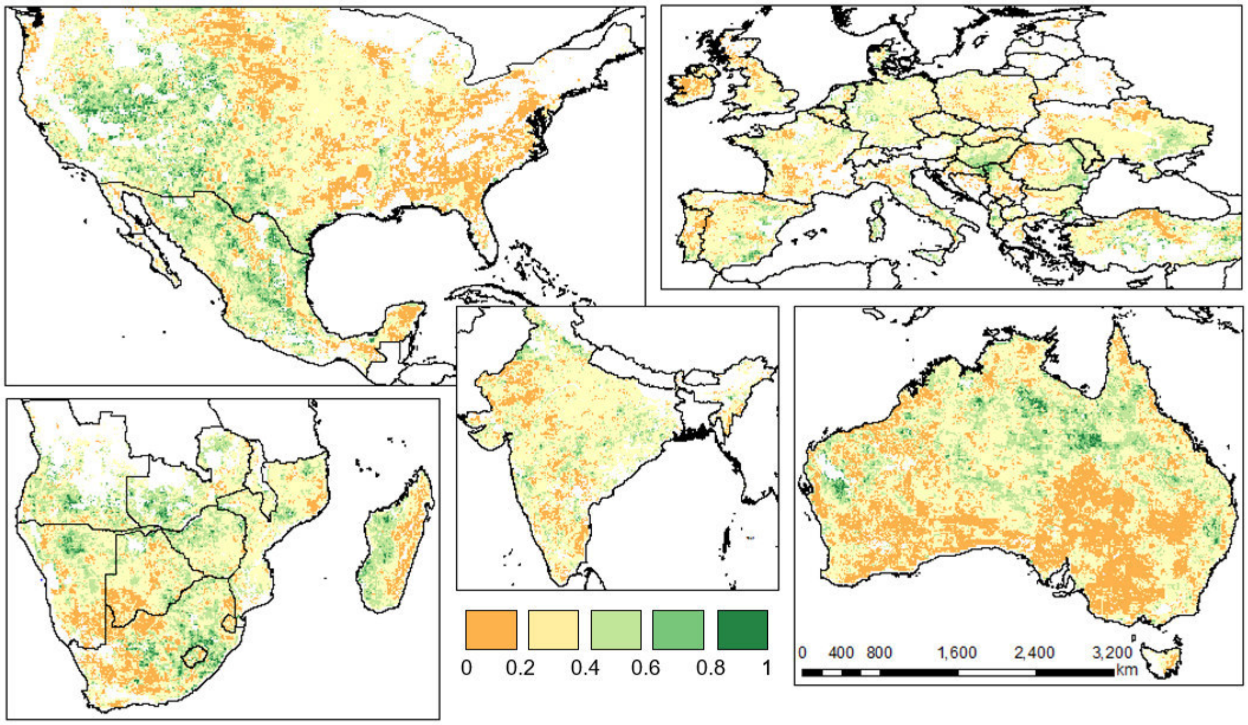

Fig. 10. Maps representing the ensemble mean weighting factor for the LST dataset according to the error maps derived from the TC analysis.

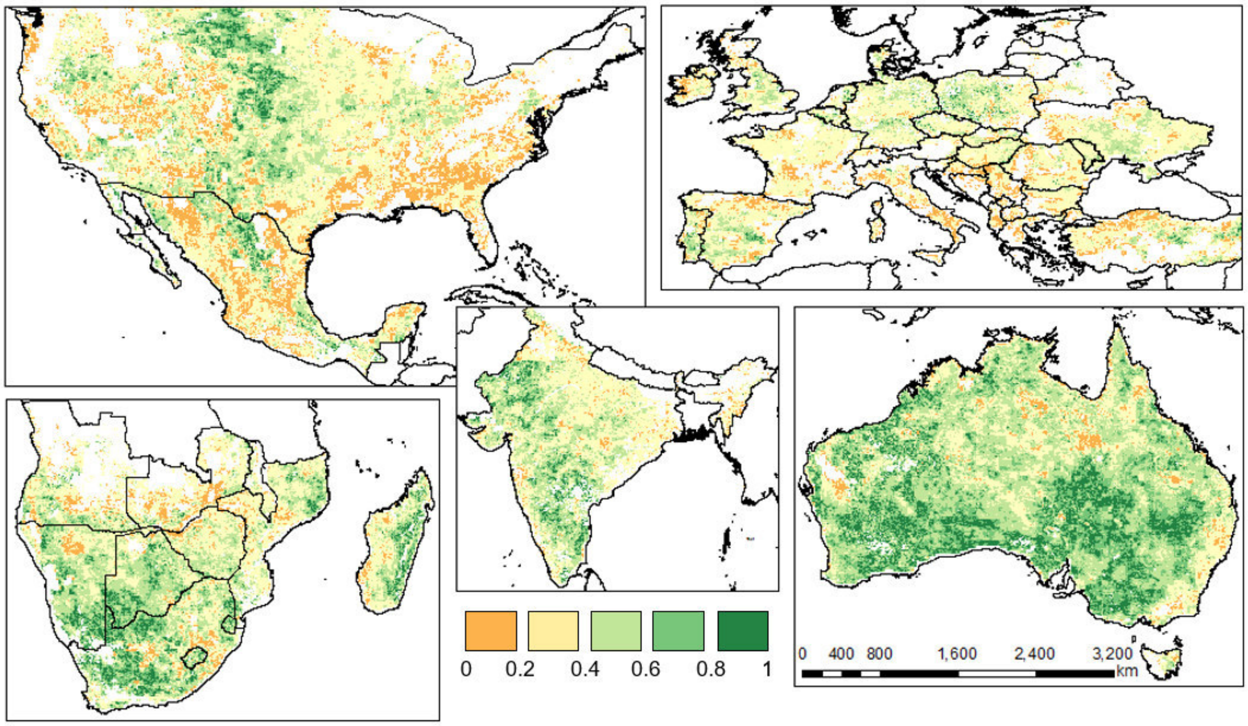

Fig. 11. Maps representing the ensemble mean weighting factor for the CCI dataset according to the error maps derived from the TC analysis.

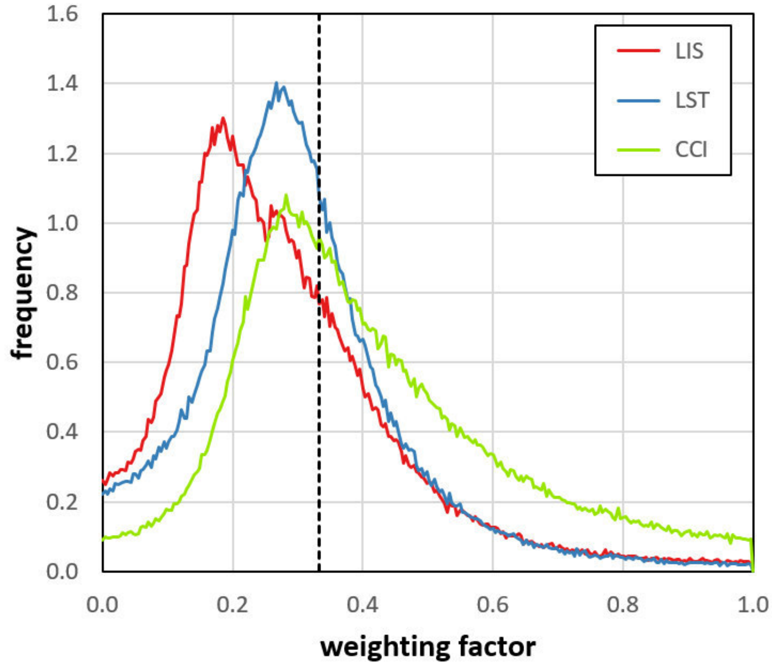

Fig. 12. Frequency distribution of the ensemble mean weighting factor for each dataset computed according to the TC analysis. The black dotted line represents the value corresponding to a simple arithmetic average (1/3).