This paper presents a Land ET product that has been generated by merging multiple ET data sets using different collocation-based approaches. While such a product would certainly be of great interest to the community, I have various major concerns about the methodology and the evaluation approach.

Response:

We sincerely thank the reviewer for your suggestions, and we have made the following changes to the manuscript according to your suggestions: (1) More detailed explanation of the merging method; (2) The performance of merged products, the inputs and result from the simple average (SA) method are comprehensively compared; (3) Modify the corresponding content of the article according to your specific suggestions; (4) The language of the article has been polished to make it more in line with the characterization of scientific papers.

(Reviewer's comments are marked in red, and our responses are marked in black)

General comments:

My biggest concern is the brute-force nature of the approach. Various collocation approaches are thrown blindly at various products with no regard given to the properties of either the products or the methods (see specific comment to L248). It seems that all possible combinations are applied and averaged, and then a selection is made (Supplement 5) without further justification or demonstration of relative performance (see below). Why selecting exactly these combinations of products and methods in these periods? What were the criteria to deem these best performing?

Much related to this comment: All the employed collocation approaches are very sensitive to error cross-correlations. While some variants tolerate/estimate cross correlation, they typically require the assumption that at least some product errors are uncorrelated. Notwithstanding, the authors seem to just apply QC to all combinations for all possible cross-correlation scenarios and then just average the results, which most likely fails terribly. This is because in all cases, cross-correlation estimates will be biased, because the assumption will be violated in either case.

A proper application of such methods would require careful consideration of the product properties. For example: If four products are considered, only two of them are allowed to exhibit non-zero error cross-correlation. QC can be applied accordingly to estimate error variances of each of the four products as well as this one error cross-covariance, but exactly which errors are correlated must be chosen a priori. Unfortunately, if more than two products exhibit correlated errors, or if the wrong data pair is assumed to have correlated errors, the whole thing breaks down and both error variance and error covariance estimates will be strongly biased. Consequently, the merging weights will also be strongly biased.

I think there's very good reason to expect strong error correlations between many products. For example, FLUXCOM is using ERA temperature for conversion, and PMLV2 uses GLDAS as an input. What about forcing data of ERA5 and GLDAS? I know that at least soil moisture simulations from ERA and GLDAS have highly correlated errors in many regions, so I don't think it will be any different for ET. Testing this could possibly be done by selecting triplets with supposedly uncorrelated errors, estimating error or variances, then replacing one product, and assessing whether the error variance estimates remain unchanged.

Response:

We are very grateful to the reviewer for your comment. For the collocation methods, the most important thing is to ensure that the errors of different products are not homologous. Therefore, we reconsidered the impact of data homology. First, in the data description section (Section 2.1), the driving data of ET products is described in more detail. Second, in the calculation, the numerical and distribution of the ECC results are analyzed, and further fusion input combinations and method selections are made based on this (in Section 4.3 and 4.4). Third, in the new Section 6, the errors of the fusion product are further discussed.

The description of the merging methodology in Sec. 3.2 is very unclear. In L281, omega is called optimal weight, even though there is never just "omega", only omega_ij, which appears to be the weight when using two data sets only. Later, in L286, omega_i is introduced as "arithmetic mean for each product", yet the equations calculate arithmetic means between weights (of data pairs), not between products. So, if I understand correctly, the authors calculate a weighted average between products, where the weights are calculated as unweighted averages between weights of data pairs that do account for error cross-correlations.

I don't know where this comes from (not from the afore-cited Kim et al. (2020), and I cannot access Bates and Granger (1969)), but I'm certain that this is not a valid least-squares solution. A least squares solution for an arbitrary number of products is provided, for example, in Eq. (2) of Gruber et al. (2019). This requires considering the cross correlation between all products in a properly constructed error covariance matrix. Is the described approach, perhaps, to account for the different possible implementations of the various method, e.g., the 30 possible options to implement quadruple collocation? For the reason stated above, I don't believe that this would be valid, and most likely does more harm than good.

Response:

We are very grateful to the reviewers for their suggestions. In the previous manuscript, there was an error in the description of the merging formula, and we corrected it accordingly (Section 3.2).

"...Given specific variances of inputs, linear combination could serve as a simple and efficient solution for data assimilation. In this study, each product (i) is assigned the optimal weight (ω_i) that minimizes the mean square error (Bates and Granger, 1969; Kim et al., 2021) using error variances ($\sigma_{\varepsilon_i}^2$) and the ECC ($\sigma_{\varepsilon_i\varepsilon_j}$) as:

$$\omega_{ij} = \frac{\sigma_{\varepsilon_i}^2 - \sigma_{\varepsilon_i\varepsilon_j}\sigma_{\varepsilon_i}\sigma_{\varepsilon_j}}{\sigma_{\varepsilon_i}^2 + \sigma_{\varepsilon_j}^2 - 2\sigma_{\varepsilon_i\varepsilon_j}\sigma_{\varepsilon_i}\sigma_{\varepsilon_j}} \quad (9)$$

The combined product θ_c is calculated as:

$$\theta_c = \sum_{i=1}^{N} \theta_i\omega_i \quad (10)$$

where ω_i is the weighted arithmetic mean for each product. For a dual-input combination, the value of ω_i is calculated as:

$$\omega_i = \frac{\omega_{ij}}{\omega_{ij} + \omega_{ji}} \quad (11)$$

For a triple-input combination, the value of ω_i is given as:

$$\omega_i = \frac{\omega_{ij} + \omega_{ik}}{(\omega_{ij} + \omega_{ik}) + (\omega_{ji} + \omega_{jk}) + (\omega_{ki} + \omega_{kj})} \qquad （12）$$

…”

This study uses the forecast combination suggested by Bates and Granger (1969), which suggests using empirical weights based on forecast variances. With further information on the error correlation, this method can work well in practice. We follow the same equation used in (Kim et al., 2021)

The issue of bias is left entirely undiscussed. The method of least squares minimizes the random error variance but doing so requires the data to be free of bias. Gruber et al. (2019) attain this by rescaling (which is only one possibility). However, as evident from e.g., Figure 11, bias is certainly present and will have a large impact on the merging. This is a problem because relative weights are calculated from random error variances and disregard biases altogether. However, when applied in the merging, they are also used to weight the biases by the same amount. Therefore, the fact that CAMELE follows FLUXNET so closely in Figure 11 appears, in my opinion, mostly serendipitous, possibly because it just so happens that - during this period - weights are evenly distributed across products. In other periods, things would look very different because in the other merging periods, much more weight is put on ERA5 (see Supplement 5). I believe this is also the reason why results appear best in the KGE, because the KGE puts a much higher weight on the contribution of bias than do the other performance metrics.

Response:

We sincerely thank the reviewer for your comment. As mentioned in the previous reply, the weight calculation method used in this study is calculated based on the information of products error and error correlation. Since the input ET products are not completely independent, the impact of ECC on the results was fully considered in the study, and the analysis and description of the ECC results was added to the new manuscript (Section 4.3). The selection of subsequent fusion products is also based on product errors and ECC calculation results (Section 4.4). In addition, we have added an analysis of the error of the fusion product to the discussion (Section 6). In the product evaluation part, in addition to the KGE coefficient, we also calculated the RMSE and R^2 coefficients, and randomly selected 4 sites (2 new sites) to compare the performance of the merged product, the input product and the weightless average (Section 5).

Related to the previous comment: The validation is insufficient and does not justify the selection of products and collocation strategies as shown in Table 2. No performance metrics are shown other than KGE statistics. How do the individual available input products perform in the different periods where data are available? How do merged products using the different collocation methods perform relative to one another, and to the performance of the individual input products? Most importantly: How would a simple unweighted average perform? For the above-described reasons, I suspect that the proposed approach cannot estimate relative weights accurately enough to outperform an unweighted average. All these aspects should be evaluated and shown separately for bias and for correlation characteristics. Least squares merging aims at improving the latter, while the largest impact appears to be in the former (which is, in fact, often found for model ensemble averages, because their bias seems to scatter rather randomly around the truth, hence averaging tends to improve that, especially in an unweighted case).

Lumping the effect of bias and correlation together in the KGE actually hampers a proper assessment of the impact of the merging algorithm.

Response:

We sincerely thank the reviewer for your comment. In addition to the KGE coefficient, we also calculated the RMSE and R^2 coefficients (new Figure 9-10), and randomly selected 4 sites (2 new sites) to compare the performance of the merged product (Figure 12-15), the input product and the weightless average (Section 5). Merged products still outperform the results by unweighted average at the site scale. We analyzed the comparison results accordingly (Section 5):

"...In addition, the results of the SA method fluctuate unreasonably at both the RU-Ha1 and CA-SF1 sites, which may be related to the error of the ERA5 product at both sites, while the performance of the merged product is much greater. This indicates that in the case of unknown product error, although the unweighted average is the optimal choice,

it may be affected by the bias of the input data and causes unreasonable estimates. Based on the product error and error correlation, a reliable weight calculation and merging method can solve this problem to some extent…”

Supplements are not referenced properly. S3 is quite unclear.

Response:

We sincerely thank the reviewer for pointing out our mistakes. We have realigned the contents of the supplements and added descriptions to the corresponding parts.

Specific comments:

L31: What about superiority / inferiority w.r.t. all the others? Why only mention one (second-best), and then only KGE?

Thanks for your comment. Since the previous description was incomplete, we haved resized the Abstract accordingly.

L33: should this be "inconsistent"?

Thanks for your comment. This sentence is related to the ET trend analysis in the previous version, which has been removed from the new manuscript.

L43: Rephrase "As the intermediate variable of soil moisture affecting air temperature"

Thanks for your comment. The introduction section has been rewritten, the previous description is incorrect, and the statement has been deleted in the new manuscript

L61: I would strongly disagree with this statement. SA is arguably the best bet if weights cannot be estimated accurately. In other words, unweighted averages often outperform badly weighted averages, and this is observed across disciplines. The authors point this out in L65.

Thanks for your comment. We are aware that our judgment of SA is wrong and only a brief description of SA is retained in the new manuscript.

L79: should be: "Su et al. (2014) proposed..." Check citation style throughout the document (The same error happens again several times in the lines that follow as well as later)

Thanks for your comment, we have adjusted the format of the citation

L83: Gruber et al. (2016) doesn't propose "quadruple collocation”, they propose collocation with an arbitrary number of n>3 data sets, referred to as extended collocation, and only demonstrate it for the case of n=4 as an example.

Thanks for your comment, we have revised the descriptions in the sections dealing with QC throughout the text

"…Gruber et al. (2016) extends the original algorithm with arbitrary number of over three data sets, and demonstrated the quadruple collocation (i.e., QC, with four data sets) as an example…"

L125: Change to "more elaborate descriptions"

Sec 2: I'd be good to be very clear about the input of all the employed models, especially to understand potential error cross-correlations. Which RS data are used for FLUXCOM?

Thanks for your comment. We have revised the description of the five ET products in Section 2.1.

L238-240: The log-transformed multiplicative error model has been preferred for precipitation products because they are assumed to exhibit a multiplicative error structure. This is not the case for other variables such as soil moisture, where the additive structure is indeed more common (and arguably more appropriate). Is there any good rationale for which to assume for the ET products used in this study?

We sincerely thank the reviewer for your comment. In section 4.2 of the new manuscript, we first analyze the applicability of additive and multiplication models to ET data. The results of Figure 5 in the new manuscript show that the multiplication model is more suitable for ET data.

L248--: This is a mere repetition of the introduction that doesn't provide any understanding of the respective methods other than how many data sets are needed. I think the readers could benefit greatly from a more thorough explanation / illustration of the differences between these approaches. What are their strengths, limitations, and assumptions? How do these relate to the properties of the products used in this study? Which would you expect to perform how? (The supplement provides mere mathematical derivations, but no insight into the properties / differences between methods.)

We sincerely thank the reviewer for pointing out our mistake. In the new manuscript, we use TC as an example to illustrate the principle of the collocation method and describe the differences between different methods (Section 3.1).

Table 2: Does this selection of products/methods during different merging periods emerge from the validation? If so, I think it'd be better to make this part of the results section alongside the validation of the different approaches. This is necessary to justify this selection.

Thanks for your comment. We have adjusted it according to your suggestions, and this part has been adjusted to after Section 4.2, and the description has been added accordingly.

L314: How's a standard deviation a validation metric? Is there any reason to believe that a low SD equates "better"? Also, no SDs are ever shown.

Thanks to the reviewer for pointing out our mistake. The SD coefficient is not used here, and it has been modified in the new manuscript.

L324: Bootstrapping cannot improve uncertainty, it can only provide confidence intervals, which is not done here.

Thanks to the reviewer for pointing out our mistake. The bootstrapping is not used here, and it has been modified in the new manuscript.

L325: How (and why, see above) was a multiplicative error model used? L331 shows additive errors

We sincerely thank the reviewer for your comment. In section 4.2 of the new manuscript, we first analyze the applicability of additive and multiplication models to ET data. The results of Figure 5 in the new manuscript show that the multiplication model is more suitable for ET data.

L328: Do you mean "Poisson distribtion"? Does ET generally follow such a distribution? (I'm not an ET expert, so I don't know). The referenced Kim et al. (2020) used a uniform distribution, but I believe that doesn't tell much anyway other than a sanity check.

Thanks to the reviewer for pointing out our mistake. The uniform distribution is used here.

Figure 2: I have the feeling there's something fishy about the synthetic experiments. For

example: Why would delta_rho increase with sample size? Isn't a lower number better, i.e., closer to the truth? Also, why should there be discontinuities in the bottom two panes?

Sincerely thanks for your comment, we rechecked this section to confirm that the results were correct. Regarding the improvement of the sample size to the result, we have two points to explain: (1) From the perspective of the effect of sample size increase on the results, the results of Figure 3 show that the effect of sample size increase is actually not large, especially when the sample size is greater than 500, $\Delta\rho$ remains stable; (2) We do not think that a small sample size will bring the data closer to the truth value, on the contrary. We haven't found relative reference to support this idea.

In addition, the previous bottom is not continuous because of a difference in picture size. The new version fixes this issue (Figure 3-4).

L428: Do the authors mean "less influenced by antecedent conditions"? This would, in fact, be a problem, because lagged TC approaches REQUIRE the variable itself to be highly auto correlated while ERRORS should be temporally uncorrelated.

Thanks to the reviewer for pointing out our mistake. This section is wrongly described, and the literature cited says that the random error of ET products has little relationship with the value of ET, and ET itself is affected by the previous situation.

Table 4: I don't understand what is shown. The description says, "Correlations against in situ", but why the columns for the different input products? And which products are being merged? All of them in all possible combinations?

We are very grateful to the reviewer for your comment. Table 4 lists the correlation coefficients between the different product errors using the Collocation methods and errors using the site measurement data. The higher the correlation between the two (closer to 1), the closer the error evaluation results of the collocation methods are to the evaluation results of the site. This is to verify that the collocation methods can be used for error evaluation of ET products.

L473: I'd recommend scaling the axis, not the values themselves.

We have adjusted according to your suggestion.

Figure 5: I don't understand what is shown. What does it mean to compare an additive

and multiplicative error’s structure? This hasn't been properly described in the Methods section. Also, the figure is very busy and hard to read. Also, what's meant with RMSE_TCA? Is TCA again used to evaluate results, after first using it for merging?

We are very grateful to the reviewer for your comment. The purpose of this section is to assess the reliability of the collocation methods by comparing the product error calculated by the collocation methods with the product error calculated on site basis. We have adjusted Section 4 and more proper description has been added accordingly.

L505: Not sure why the results section starts here... A lot of results are already shown before that.

This section shows the results of the evaluation of the fusion product, which is different from Section 4. We changed the name of this section to “Product Evaluation”.

Figure 12 is the same as Figure 11 (caption seems to be correct, but the figure seems to be wrong).

Thanks to the reviewers for their comments. This section has been adjusted, we have fixed previous bugs, and the results of the comparison of the two sites have been added. In addition, the results of the SA method are also shown in the figure.

Figures 13-14: Hard to compare visually... Would it make sense to show difference maps?

Figures 15-16: How confident are you that trends aren't introduced by the merging algorithm? I understand from Table 2 that different products/methods are used in different periods. This could introduce trends just by having jumps in the data, especially if no bias correction is applied (see general comments).

Thanks for your comment, the global trend analysis of ET is not the focus of this article, so this section has been removed in the new manuscript. The new Section 6 focus on the discussion of merged products.

Data repository:

The description should be a description of the data to make it easier for people to understand/use them, not a mere copy of the abstract of the manuscript.

We sincerely thank the reviewer for pointing out the mistake. We have updated the description on Zenodo:

“This dataset provides a long-period estimation of global land evapotranspiration over two resolutions: (1) 0.1°-8day-average resolution, covering 2001-01-01 to 2019-08-29; (2) 0.25°-Daily resolution, covering 1981-01-01 to 2020-12-31. The product was merged using a collocation-based approach. The inputs included: (1) ERA5-land-hourly total evaporation products; (2) Penman-Monteith-Leuning Evapotranspiration V2 (PMLV2); (3) The FluxCom-RS ensemble of global energy flux; (4) the Global Land Evaporation Amsterdam Model version 3.3a (GLEAMv3.3a); (5) The Global Land Data Assimilation System-Noah model (GLDASv2.1 Noah). Five collocation algorithms were used for characterizing the uncertainties of inputs. And the optimal weights of each product were calculated by minimizing the mean square error based on the error information. The CAMELE product was further evaluated against 82 FLUXNET sites and showed average R2 value of 0.73 and 0.68 over 0.1° and 0.25° resolutions, respectively. CAMELE ET can be used for hydrological studies and regional investigation of water resources management, etc. For further information, please check our related publication.

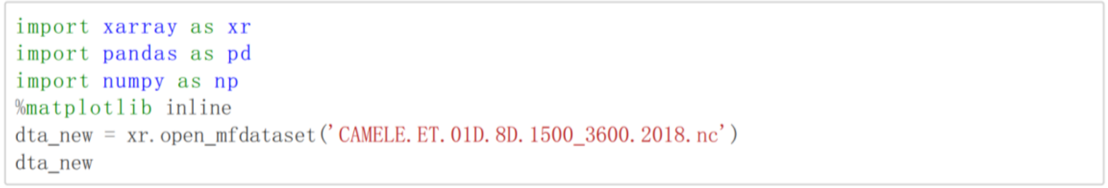
Section A : 0.1°-8day-average resolution from 2001 to 2019. (10.5281/zenodo.6616791)

Section B: 0.25°-daily resolution from 1981 to 2020. (10.5281/zenodo.6616815)”.

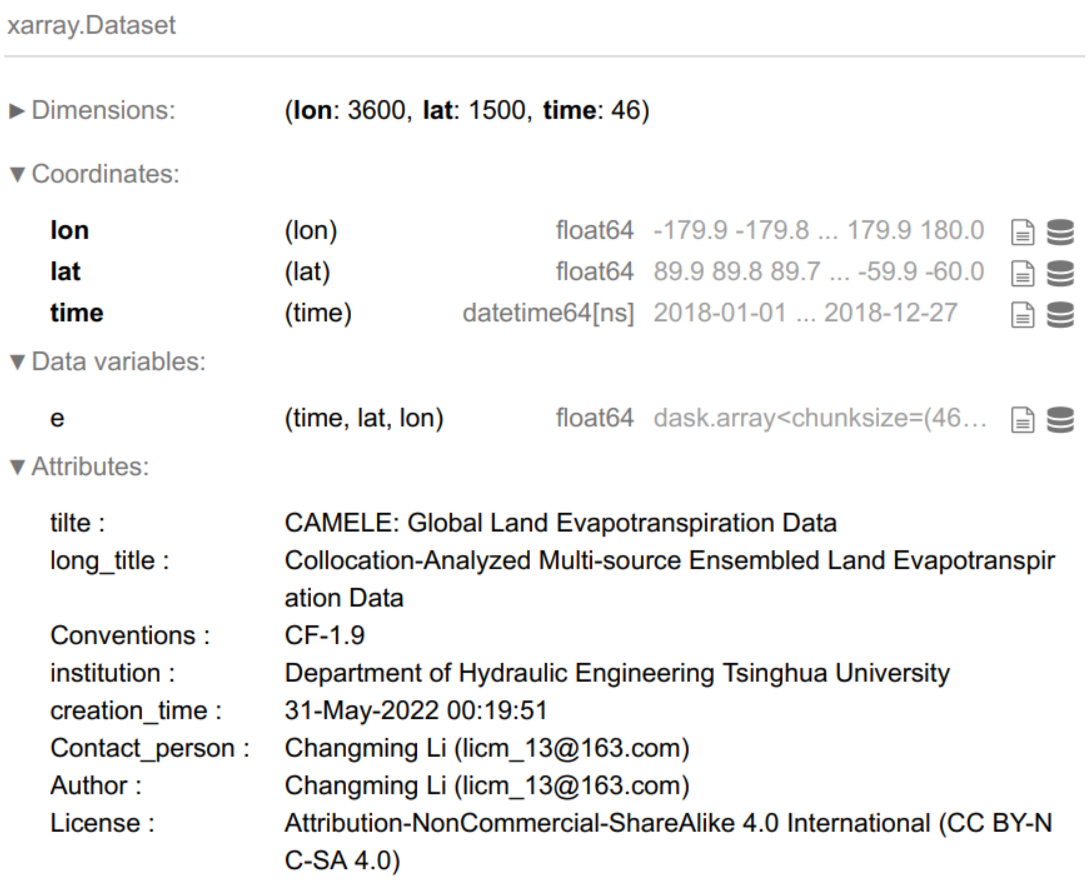
I couldn't open the data in panoply because "Axis includes NaN value(s)". This seems to be the case for all 3 dimensions. Please fix the data files so that dimensions include only valid data.

We sincerely thank the reviewer for your comment. We have checked our dataset using Python and it worked fine. Randomly select two files as example:

(1) Over 0.1° resolution

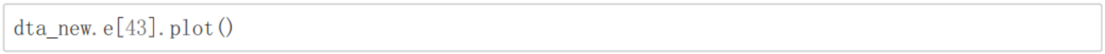


```
import xarray as xr
import pandas as pd
import numpy as np
%matplotlib inline
dta_new = xr.open_mfdataset('CAMELE.ET.01D.8D.1500_3600.2018.nc')
dta_new
```



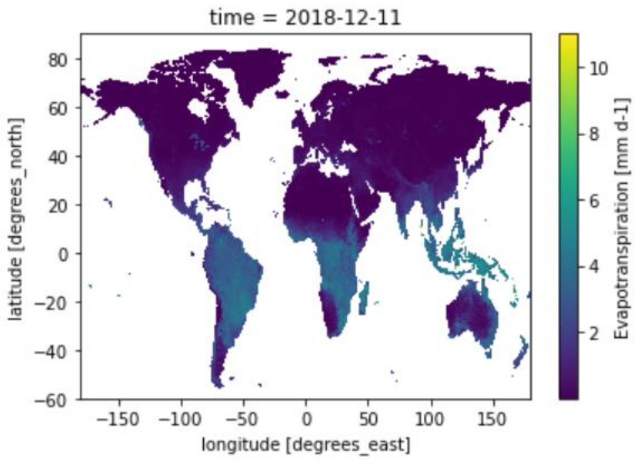
```
xarray.Dataset

▶ Dimensions:        (lon: 3600, lat: 1500, time: 46)
▼ Coordinates:
    lon              (lon)                  float64  -179.9 -179.8 ... 179.9 180.0
    lat              (lat)                  float64  89.9 89.8 89.7 ... -59.9 -60.0
    time             (time)           datetime64[ns]  2018-01-01 ... 2018-12-27
▼ Data variables:
    e                (time, lat, lon)       float64  dask.array<chunksize=(46…
▼ Attributes:
    tilte :          CAMELE: Global Land Evapotranspiration Data
    long_title :     Collocation-Analyzed Multi-source Ensembled Land Evapotranspir
                     ation Data
    Conventions :    CF-1.9
    institution :    Department of Hydraulic Engineering Tsinghua University
    creation_time :  31-May-2022 00:19:51
    Contact_person : Changming Li (licm_13@163.com)
    Author :         Changming Li (licm_13@163.com)
    License :        Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-N
                     C-SA 4.0)
```

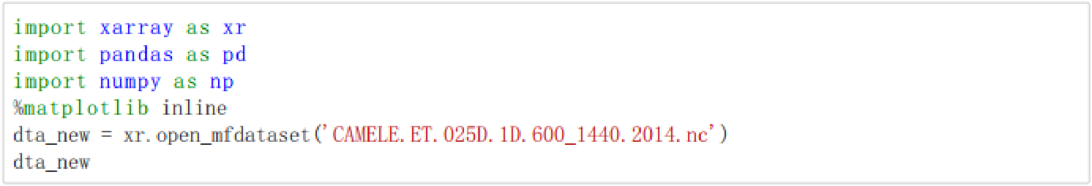


```
dta_new.e[43].plot()
```

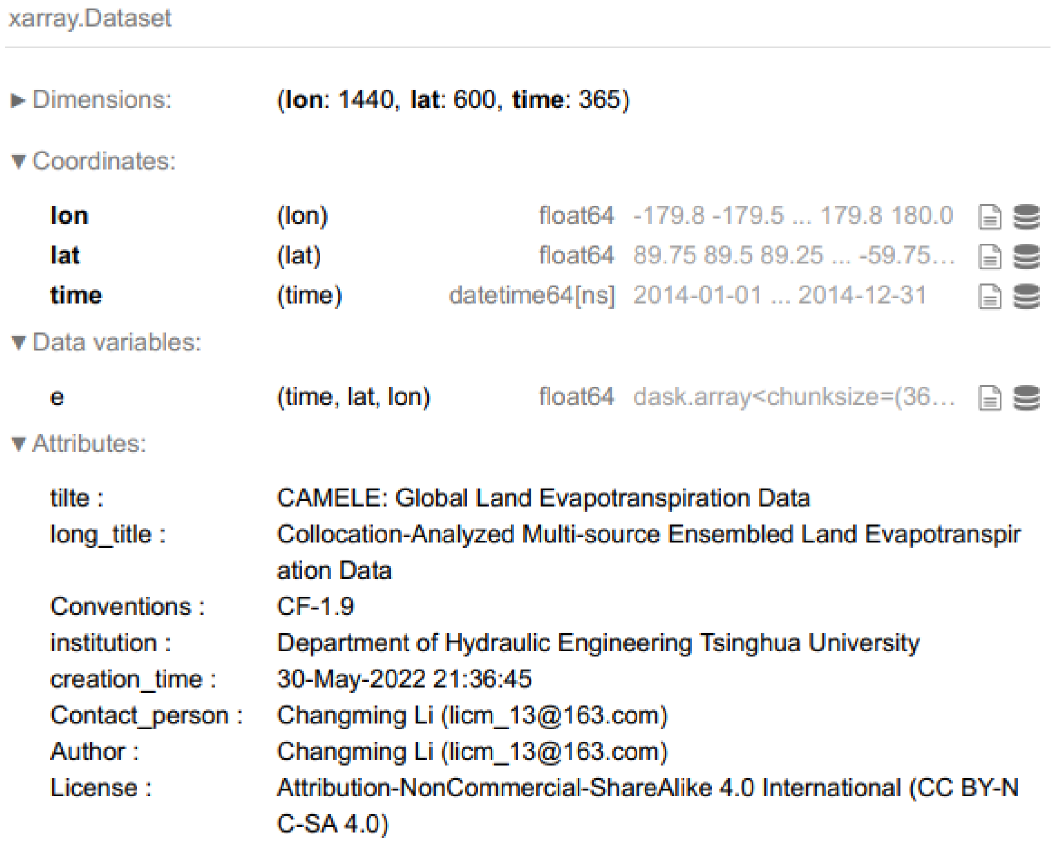
```
<matplotlib.collections.QuadMesh at 0x1ee63a22280>
```



(2) Over 0.25° resolution



```
import xarray as xr
import pandas as pd
import numpy as np
%matplotlib inline
dta_new = xr.open_mfdataset('CAMELE.ET.025D.1D.600_1440.2014.nc')
dta_new
```



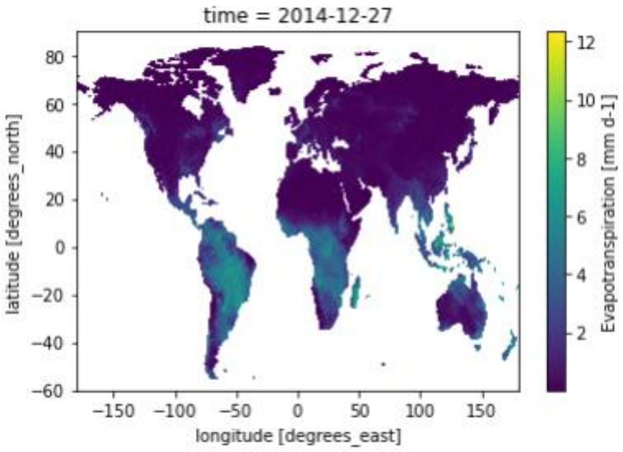
```
xarray.Dataset

▶ Dimensions:          (lon: 1440, lat: 600, time: 365)

▼ Coordinates:
    lon               (lon)                  float64   -179.8 -179.5 ... 179.8 180.0
    lat               (lat)                  float64   89.75 89.5 89.25 ... -59.75...
    time              (time)          datetime64[ns]   2014-01-01 ... 2014-12-31
▼ Data variables:
    e                 (time, lat, lon)       float64   dask.array<chunksize=(36...
▼ Attributes:
    tilte :            CAMELE: Global Land Evapotranspiration Data
    long_title :       Collocation-Analyzed Multi-source Ensembled Land Evapotranspir
                       ation Data
    Conventions :      CF-1.9
    institution :      Department of Hydraulic Engineering Tsinghua University
    creation_time :    30-May-2022 21:36:45
    Contact_person :   Changming Li (licm_13@163.com)
    Author :           Changming Li (licm_13@163.com)
    License :          Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-N
                       C-SA 4.0)
```

```
dta_new.e[360].plot()
```

```
<matplotlib.collections.QuadMesh at 0x1ee04d7ba00>
```


time = 2014-12-27
80
60
40
20
0
−20
−40
−60
latitude [degrees_north]
−150
−100
−50
0
50
100
150
longitude [degrees_east]
12
10
8
6
4
2
Evapotranspiration [mm d-1]