Interactive comment on "Historical gridded reconstruction of potential evapotranspiration for the UK" *by* Maliko Tanguy et al.

The authors' first response to reviewers: Anonymous Referee #1

The reviewer's comments are in black, and our response is in blue.

We would like to thank Referee #1 for his/her valuable and constructive comments which will help improve the manuscript. Below is our point-to-point response (in blue).

Historical gridded reconstruction of potential evapotranspiration for the UK Tanguy et al.

This paper produces a gridded reconstruction (at daily and monthly timesteps) of PET for the UK (excluding Northern Ireland) for the period 1891-2015. The paper presents the selection of methods and decisions in producing the final dataset and assesses the performance of the selected approaches relative to a naive climatology and CHESSPET as a surrogate for observations. The work produces a valuable dataset of practical utility, especially for river flow reconstructions. Overall the authors do a good job and I support publication. However I have a number of specific points that I wish the authors to consider before publication. These relate to both the underpinning science and uncertainties but also the presentation of the paper to help readers interpret what was done in a clearer way.

We would like to clarify that the new PET dataset **does** cover Northern Ireland (NI). It is the high resolution CHESS PET dataset (available for 1961-2015) that is not available for NI. As CHESS PET was used for the assessment, the performance metrics could not be calculated for NI, but the data was produced, and we expect performance to be similar due to geographic proximity. We will revise the text to make sure this is clear.

Specific Comment

The work uses UKCP09 monthly temperature (1910-2015), together with a gridded dataset of monthly temperature from the historical drought project. I am left wondering about some details of the underlying temperature datasets – i) is there a decrease in station density underpinning these products the further back in time one goes? ii) does this affect the spatial distribution or errors in your dataset? iii) does the joining of both datasets create a break in the data, has this been checked? iv) has the underlying temperature data been homogenised? If so, how, if not what might this mean for the derived product here. So far as I can see these issues are not outlined or discussed in the paper. A fuller discussion on the uncertainties associated with the derived product is required in the discussion. You indicate that MAPE can be used to estimate uncertainties but only based on the selected method. There are other, potentially greater uncertainties that should be transparently laid out.

As pointed out by Referee #1, we have only investigated the uncertainties coming from the method used to calculate PET, but not from the underlying temperature dataset used. We will add a paragraph in the manuscript to discuss the use of the underlying temperature data, and its implication on the final uncertainty. We will also more fully discuss all these issues in the supplementary information.

More specifically:

i) There is indeed a change in station density underpinning the temperature grids. According to information provided by the Met Office, the station density gradually increased from 74 stations across the country in 1891 to a peak of 672 in the mid-1990s, after which it decreased again to reach a total of 355 stations in 2015.

ii) Legg (2015) has investigated extensively the effect of network density on the error in gridded dataset in the UK, and his results suggest that the change in density observed here would only lead to a minor increase in error in temperature. An increase in the root-mean-square error of less than 0.2°C is observed for most cases when the network density changes from 570 to 75 stations across the UK. This reflects the spatial coherence in the temperature data.
We have conducted a sensitivity analysis of PET on errors in input temperature, when using McGuinness-Bordne equation. We have found that a +/- 0.2°C in input temperature translates into a 0.5% to 2% difference (with an average of 0.8%) in PET estimation. We consider these differences negligible in comparison to the uncertainties arising from the PET method itself.

iii) The joining of both datasets does not create a break in the dataset, as the exact same methodology was applied to derive the grids (method described in Perry and Hollis, 2005). The dataset previously only existed from 1910, as the density of stations was too sparse prior to that. However, within Historic Droughts project, historic data has been rescued and digitised by the Met Office, which has allowed them to extend the gridded temperature back to 1891.

iv) The gridded temperature product used in this study is a standard national product produced by the UK Met Office National Climate Information Centre (NCIC). All the underlying stations are part of the climatological network administered by the Met Office, which are subject to common observation methods, regular site inspections and instrument calibration. The data is also subject to quality control procedures prior to archiving. For all this, Prior and Perry (2014) concluded that, even if individual station records were not tested for homogeneity, the effect of inhomogeneities in the gridded product is considered minimal. The interpolation and regression methods used to create the gridded product, which takes into account factors such as latitude and longitude, altitude and terrain shape, coastal influence, and urban land use, reduces the impact of station openings and closures on homogeneity, although it can't be removed entirely, especially in areas of complex topography or sparse station coverage (Perry and Hollis, 2005).

References:

Legg, T. (2015), Uncertainties in gridded area-average monthly temperature, precipitation and sunshine for the United Kingdom. Int. J. Climatol., 35: 1367–1378. doi:10.1002/joc.4062

Prior, M. J. and Perry, M. C. (2014), Analyses of trends in air temperature in the United Kingdom using gridded data series from 1910 to 2011. Int. J. Climatol., 34: 3766-3779. doi:10.1002/joc.3944

While the authors do outline the different temperature based PET methods in a table it would be beneficial to include a discussion of the main differences between each method in the text. Are these an exhaustive selection, if not, why these methods, why not others.

We will add some text to highlight the main differences between each method in the manuscript. The physical basis for estimating evaporation using temperature alone is that both terms of the combination equation (the energy required to sustain evaporation and the energy removed from the surface as water vapour) are generally related to temperature (Shuttleworth, 1993).

The main difference between the different temperature-based formulations, lies in the way temperature is linked to PET to simulate the effect of the full set of variables normally required in the combination equations. Most temperature-based equations use day length or related variables (Hamon, 1961; Blaney and Criddle, 1950; Kharrufa, 1985; MOHYSE: Fortin, 2006 and Thornwaite, 1948), except McGuinness and Bordne (1972), and the derived Oudin (2005)'s equation which use extraterrestrial radiation instead. Blaney-Criddle equation has also an additional parameter k, which depends on crop type. Most of these equations were developed for the USA, except MOHYSE (which was developed in Quebec), Kharrufa (developed for arid regions) and Oudin (developed in Australia, USA and France).

We have tested all widely used temperature-based PET equations that we were aware of, using mean temperature data as the only climatic variable.

There are a lot of datasets/methods/calibration designs/verification methods used in the paper and at times it is hard to follow. Some effort at making the presentation clearer through signposting is necessary. The authors do include a work flow diagram but this too is complex. Perhaps a table describing the different datasets developed and why used would be useful. Perhaps the flow diagram could be split in two – separate for validation with some further detail to help interpretation in both parts.

We realise the paper can be difficult to follow due to multiple datasets/methods/assessments. We will make sure there will be more signposting in the revised manuscript to help the readers' comprehension. We will add the following table in the supplementary information (new table A3) to summarise datasets used, and we will replace the current flow diagram (Figure 1) with the following revised version (new Figure 1). The text in the manuscript will also be modified to better follow the Stages described in the new figure.

Table A3: Summary of temperature and PET datasets used in this study a) temperature data to investigate effect of temporal distribution of data on the output PET estimation, b) temperature data to investigate effect of spatial resolution of the data on the output PET estimation, and c) PET data used to calibrate and evaluate equation and final PET output.

a) Temperature datasets used as input data for temperature-based PET equations. Multiple versions were used to investigate the effect of temporal distribution of the data on output PET estimation

Dataset short name	Resolution	Description	Comment
CHESS-temp daily	1 km x 1 km Daily	CHESS-met high resolution mean daily temperature: Part of a larger dataset developed by CEH for environment modelling applications, available for 1961-2015	"Best" available gridded daily temperature data for Great Britain
CHESS-temp clim	1 km x 1 km Daily	CHESS daily mean temperature climatology: Long term average (1961-1990) of daily mean temperature, derived from CHESS-temp daily	Default option that could be used even if no temperature data were available
CHESS-temp monthly I	1 km x 1 km Monthly disaggregated to daily	CHESS daily mean temperature derived from monthly averages, constant during the month. Step changes in temperature between consecutive months	To investigate whether temporal disaggregation method (from monthly to daily) has an effect on output PET estimation
CHESS-temp monthly II	1 km x 1 km Monthly disaggregated to daily	CHESS daily mean temperature derived from monthly averages, interpolated using pchip	
CHESS-temp monthly III	1 km x 1 km Monthly disaggregated to daily	CHESS daily mean temperature derived from monthly averages, disaggregated to daily using CHESS daily mean temperature climatology pattern	

b) Temperature datasets used to assess spatial resolution for the best performing PET method

Dataset short name	Resolution	Description	Comment
UKCP09-temp monthly I	5 km x 5 km Monthly disaggregated to daily	UKCP09 daily mean temperature derived from monthly averages, constant during the month	Two temporal disaggregation methods tested.
UKCP09-temp monthly II	5 km x 5 km Monthly disaggregated to daily	UKCP09 daily mean temperature derived from monthly averages, interpolated using pchip	

c) PET datasets used to calibrate the equations and assess the output PET

Dataset short name	Resolution	Description	Use
CHESS-PM	1 km x 1 km Daily and monthly	CHESS-PET 1-km grids, daily (and monthly) time series available for 1961-2015, calculated using the Penman-Monteith (PM) equation for FAO-defined well-watered grass	1) calibration of the temperature-based PET equations (1961-1990) 2) Evaluation of the equations (1991-2012) 3) Evaluation of the final gridded product (1991-2012)
CHESS-PM climatology	1 km x 1 km Daily and monthly	Daily (and monthly) PET long term average, calculated from CHESS-PM for 1961 to 1990	used as a 'naïve method' against which the PET reconstruction methodology can be tested to assess performance

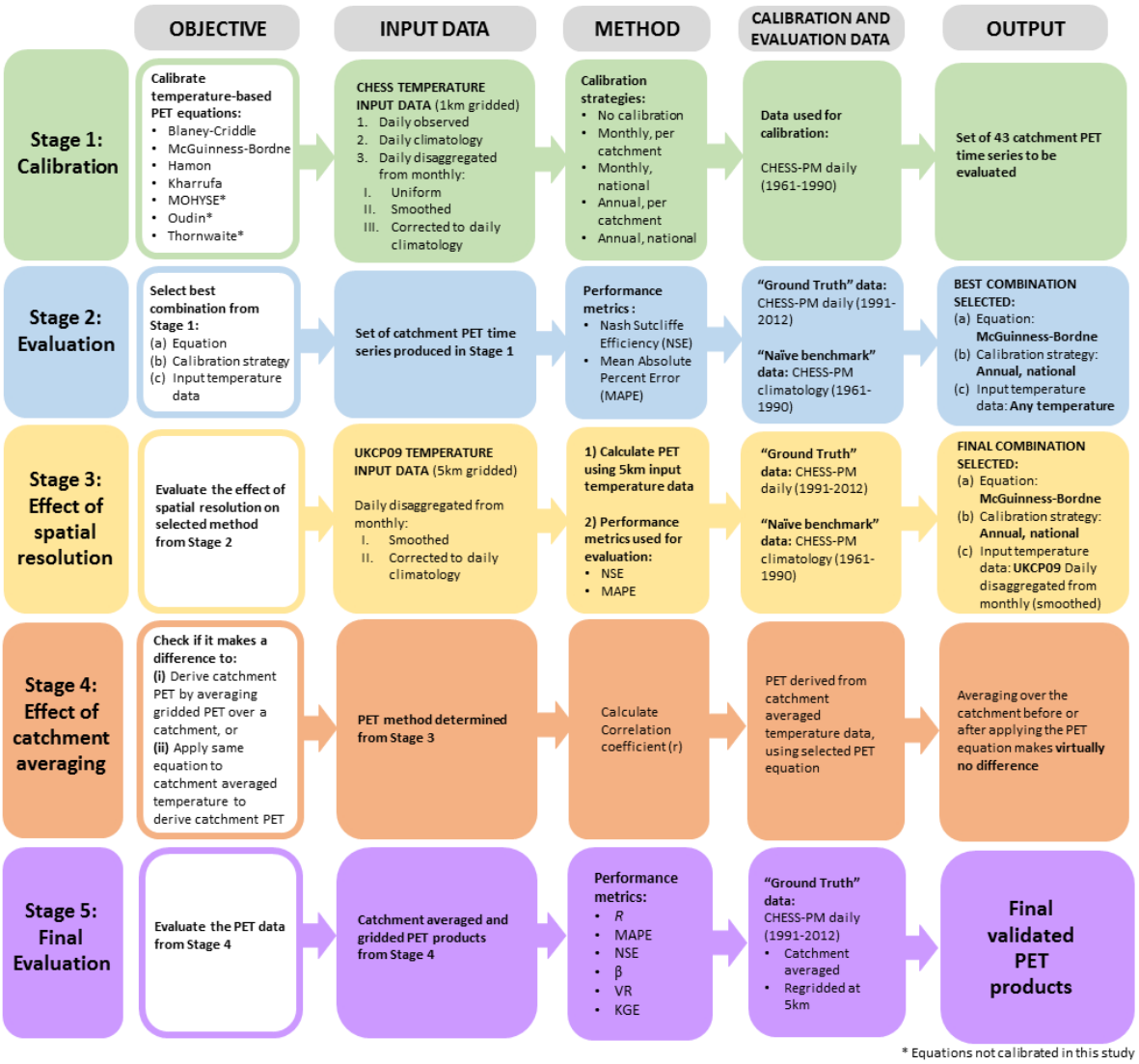

Figure 1: Work flow diagram of the evaluation procedure of the PET equations and final PET gridded product. This process was made in five stages: in stage 1, the equations were calibrated using different calibration strategies and different input temperature data; in stage 2, the multiple combinations of PET equation/calibration approach/temperature input data were evaluated; in stage 3, the effect of spatial resolution of the input temperature data was assessed. These three first stages led to the selection of PET equation, calibration strategy and input dataset used to produce the final gridded PET product. In a fourth stage, the effect of calibrating the equations at catchment scale was investigated; and finally, in stage 5, a final evaluation of the new gridded PET product was carried out both at catchment-scale and at grid-scale. Stages 1, 2 and 3 used the set of 43 catchments shown in Fig. 2a, whereas stages 4 and 5 used the full set of 306 evaluation catchments shown in Fig. 2b.

I do not know what the pchip method is or how it performed, or why is was used above other approaches for disaggregation. Would use of another methods affect results? Pchip is mentioned in the abstract and once or twice in the paper but we have not details about its application.

Pchip stands for Piecewise Cubic Hermite Interpolating Polynomial, which is an interpolation method in which a cubic polynomial approximation is assumed over each subinterval.

The results (Fig. 4 and table A1 in supplementary material) show that the daily temporal distribution of temperature over the month has a minimal effect on the PET outputs. Therefore choosing pchip or a different interpolation method for the temporal disaggregation would only have a marginal effect on the estimated PET.

The following text will be added to the manuscript:

“Pchip stands for Piecewise Cubic Hermite Interpolating Polynomial, which is an interpolation method in which a cubic polynomial approximation is assumed over each subinterval. Arandiga et al. (2016) describe this interpolation scheme in detail together with its advantages, mainly that it is both accurate (preserves values at the nodes) and preserves monotonicity. Pchip was selected for the present study because (i) the fitted curve passes through observed values at inflexion points unlike spline or quadratic methods, for example, and (ii) it does not require re-fitting when the period of application is extended as each subinterval is treated separately.”

Reference: F. Aràndiga, R. Donat, M. Santágueda, 2016, The PCHIP subdivision scheme, Applied Mathematics and Computation, Volume 272, Part 1, Pages 28-40, ISSN 0096-3003, https://doi.org/10.1016/j.amc.2015.07.071.

Minor Points/Technical corrections

Is there any influence of catchment size on the results?

Fig.R1 (daily PET) and Fig.R2 (monthly PET) below show the performance metrics vs the catchment area (logged scale for x axis as there are many small catchments and only a few very large ones).

Skill for monthly PET is higher than for daily PET for all size catchments. However, no clear relationship between the performance metrics and the catchment size stands out. The performance depends more on location (North vs South, near the coast vs inland) as discussed in section 4.2 and shown in Fig. 5 and 6, than on catchment size.

We will add a sentence in the manuscript saying that we have tested the relationship between performance and catchment area, but that no clear relation was found.

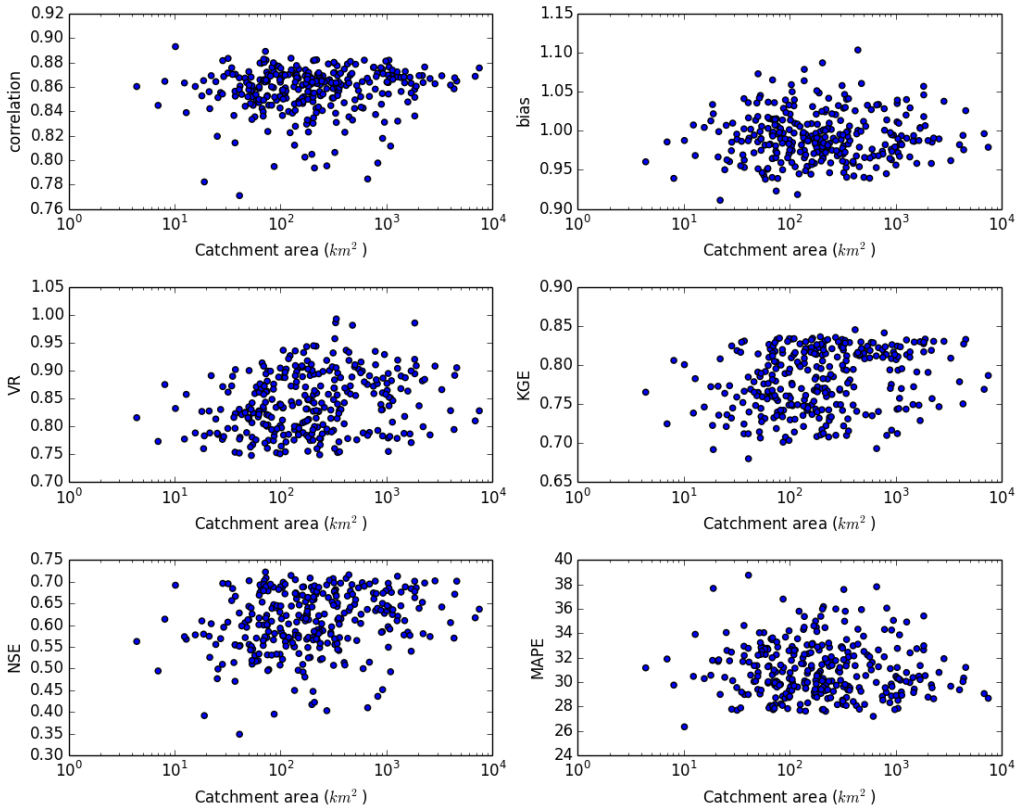

Fig. R1: Relationship between catchment size and performance metrics for daily PET data.

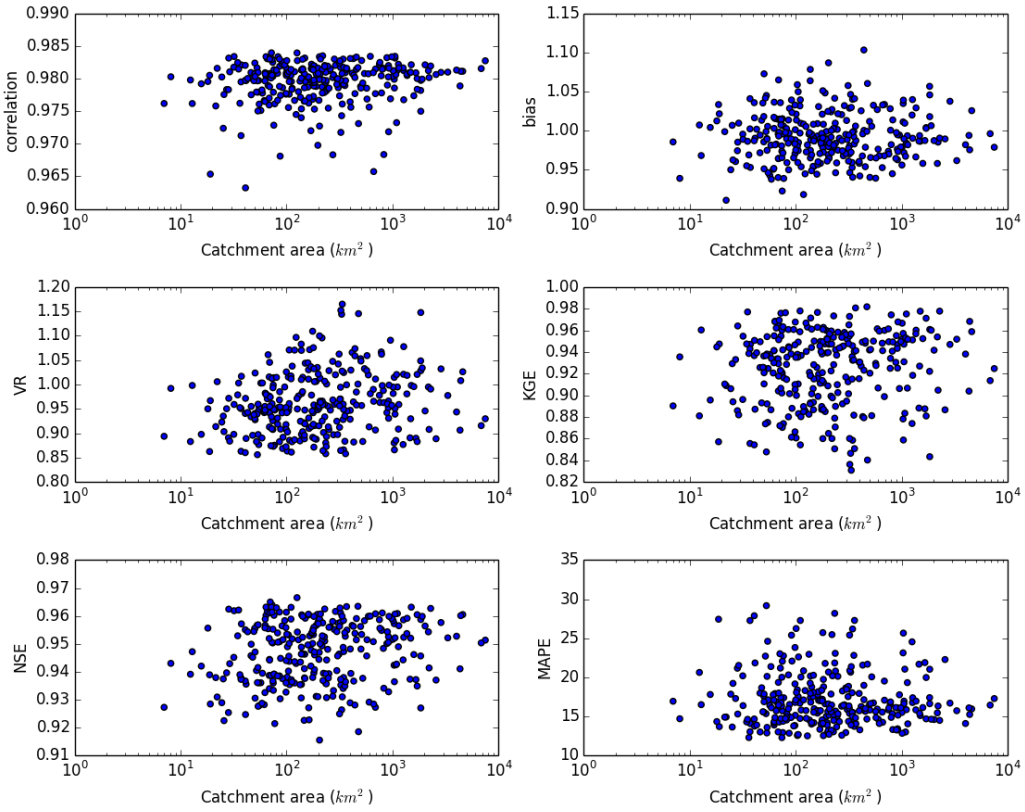

Fig. R2: Relationship between catchment size and performance metrics for monthly PET data.

In terms of extending to Northern Ireland, could reanalysis data be used here in future work. I think there is an onus on the authors to discuss how currently limitations may be overcome in future work.

As mentioned earlier, Northern Ireland is included in the new PET dataset. We will make sure this is clearer in the revised manuscript.

Regarding the use of re-analysis data, it could indeed be an alternative way to estimate past PET. The objective of the current work was to produce the best possible PET data given the available observed data, but the possibility to use reanalysis data as an alternative or complementary source will be discussed in the revised manuscript, and is actually being explored in further work.

Abstract – needs to be reworked. You state that PET is needed at daily or shorter time step – do you mean longer? You examine monthly and daily, not hourly. I suggesting removing the word 'reconstructing' from line 3 of abstract – application of models before 1960 could be for a number of purposes – much flow data commences before PET data. Sentence commencing line 15 is too long, needs to be broken into at least two sentences. You need to tell the reader in the abstract what naive methods are. Line 25 -27 is perhaps too detailed for an abstract. What is pchip? Abstract also needs a final statement on envisaged uses of the dataset.

We have reworked the abstract, taking into account all of your suggestions. In particular, we have removed the mention that PET is needed at daily or 'shorter time-step' as we agree that it is confusing. What we meant is that for hydrological modelling, PET is usually needed at daily or shorter time-steps (sub-daily, hourly), the shorter time-steps mostly needed for flood studies. However, as we haven't examined sub-daily PET here, the latter part is irrelevant, and therefore removed from the abstract. We also removed the mention to pchip interpolation method, as this is very specific and is not relevant in the abstract. The revised abstract is shown below.

Abstract. Potential Evapotranspiration (PET) is a necessary input data for most hydrological models and is often needed at a daily time-step. An accurate estimation of PET requires many input climate variables which are in most cases not available prior to the 1960s for the UK, nor indeed most parts of the world. Therefore, when applying hydrological models to earlier periods, modellers have to rely on PET estimation derived from simplified methods. Given that only monthly observed temperature data is readily available for the late 19th and early 20th century at a national scale for the UK, the objective of this work was to derive the best possible UK-wide gridded PET dataset with the limited data available.

To that end, firstly, a combination of (i) seven temperature-based PET equations, (ii) four different calibration approaches and (iii) seven input temperature data were evaluated. For this evaluation, a gridded daily PET product based on the physically-based Penman-Monteith equation (the CHESS PET dataset) was used, the rationale being that this provides a reliable 'ground-truth' PET dataset for evaluation purposes, given that no directly observed, distributed PET datasets exist. The performance of the models was also compared to a 'naïve method', which is defined as the simplest possible estimation of PET in the absence of any available climate data. The 'naïve method' used in this study is the CHESS PET daily long term average (the period from 1961 to 1990 was chosen), or CHESS-PET daily climatology.

The analysis revealed that the type of calibration and the input temperature dataset had only a minor effect in the accuracy of the PET estimations at catchment scale. From the seven equations tested, only the calibrated version of the McGuinness-Bordne equation was able to outperform the 'naïve method' and was therefore used to derive the gridded, reconstructed dataset. The equation was calibrated using 43 catchments across Great Britain.

The dataset produced is a 5-km gridded PET dataset for the period 1891 to 2015, using as input data for the PET equation the Met Office 5-km monthly gridded temperature data available for that time period. The dataset includes daily and monthly PET grids and is complemented with a suite of mapped performance metrics to help users assess the quality of the data spatially.

This dataset is expected to be particularly valuable as input to hydrological models for any catchment in the UK.

The data can be accessed here: https://doi.org/10.5285/17b9c4f7-1c30-4b6f-b2fe-f7780159939c.

Introduction Is there a word missing from end of the first sentence?

We will rephrase to: Potential evapotranspiration is a conceptual variable which measures the atmospheric demand for moisture from open surface water.

Line 7 – suggest the word approaches rather than formulations. Also what are combination methods – used without any explanation in the first instance.

'formulations' will be replaced by 'approaches'.

The sentence about combination methods will be rephrased as: "The most complex are based on physical processes accounting for the energy available to a plant to evaporate during photosynthesis, and the amount of water that can be dissipated in the atmosphere (Penman, 1948;Monteith, 1965). They are referred to as combination methods as they combine the energy balance with the mass transfer method."

Line 13 – temperature as 'a' proxy

Will be changed

Page 3, line 1 – do you have a reference to support the claim that PET is mainly used for hydro modelling. If not suggest widely used rather than mainly.

'Mainly' will be replaced by 'widely'

Line 9 – as 'an' alternative

Will be changed

Line 12 – longest period in the UK – give us some context about the historical length of observations. Also maybe suggest that temp and precip are 'among' those with longest records. Other variables such as Sea Level Pressure also have been rescued historically.

Detailed climatic variables are available in the UK in a high resolution gridded format (5km) from 1961 onwards. Sunshine hours is available from 1929, temperature and precipitation from the Met Office was available from 1910, but has recently been extended back in time thanks to historical data rescuing effort by the Met Office funded by Historic Droughts project. Monthly temperature gridded data was available to project partners from 1891. Sea Level Pressure data is also available from late 19th century (Met Office HADSLP2 product), however the spatial resolution is much coarser (5 degrees).

Some text will be added to the manuscript to give some context about the historical length of observations.

Actually, thinking about this, could reanalysis products that assimilate SLP observations be used to supplement this work and temporally extend the record in future work? Perhaps you could come to this potential (or not) in the discussion.

As mentioned earlier, the use of reanalysis data as an alternative or complementary method is certainly very valuable, and is being explored in further work. This will be discussed in the revised manuscript.

Line 13 – where only temperature 'data' are available.

Will be added

Line 24 – for the international reader use of Great Britain, UK, NI can be confusing. Keep the terminology the same – suggest UK (incl. NI).

Noted. We will avoid the use of Great Britain, and replace by terminology UK (incl. NI) and UK (excl. NI)

When introducing temperature data please include in the bracket the term henceforth. . ..eg. (henceforth CHESS-temp daily)

Will be added

Page 4 line 8 – Because of the coarser temporal and spatial resolution of temperature data prior to 1961 – please give us some details on this and the PET dataset is dependent on this data.

We will add some explanation here: 'Prior to 1961, temperature data is only available at a 5km spatial resolution and monthly time-step.'

Line 10-25 – you need to help the reader more in introducing these datasets and study design. The text is a little terse here and too brief. Some further reasoning and justification required. Eg. above, line 23 (iii) – I am not sure exactly what is going on here.

We will add some explanation to help the reader better understand this section. We have also added a table (shown in page 3 of this document) to summarise all the datasets used in this study.

Line 9-27 (page 4) will be replaced by the following text, which hopefully helps the reader better understanding these various datasets:

'Prior to 1961, temperature data is only available at a 5km spatial resolution and monthly time-step. Because of this coarser temporal and spatial resolution of temperature data in the earlier period, alternative datasets were generated and used in the analysis to quantify the sensitivity of PET derivation to temperature input, and are summarised in table A3(a and b) in the supplementary information:

- CHESS daily mean temperature climatology (1-km grids) (CHESS-temp clim): long term average (1961-1990) of daily mean temperature, derived from CHESS-temp daily. This provides a default option that could be used even if no temperature data were available in the past (or future). This gives a day-to-day variability pattern of temperature throughout the year, which is then repeated every year.
- CHESS daily mean temperature derived from monthly averages (1km grids). Different methods to disaggregate monthly temperature into daily data were tested:
  - (i) Constant temperature during the month (CHESS-temp monthly I). This means there are step changes in temperature between consecutive months.
  - (ii) Interpolated using pchip (piecewise cubic hermite interpolating polynomial) method for a smooth transition between months (CHESS-temp monthly II). Pchip stands for Piecewise Cubic Hermite Interpolating Polynomial, which is an interpolation method in which a cubic polynomial approximation is assumed over each subinterval. Arandiga et al. (2016) describe this interpolation scheme in detail together with its advantages, mainly that it is both accurate (preserves values at the nodes) and preserves monotonicity. Pchip was selected for the present study because (i) the fitted curve passes through observed values at inflexion points unlike spline or quadratic methods, for example, and (ii) it does not require re-fitting when the period of application is extended as each subinterval is treated separately.
  - (iii) Disaggregated to daily using CHESS daily mean temperature climatology pattern (CHESS-temp monthly III). The daily relative variation in temperature follows the climatology, but for each month, the daily values are adjusted so that monthly mean temperatures are correct. In other words, CHESS daily climatology data is shifted uniformly so the monthly mean temperature matches the CHESS monthly temperature data.

- UKCP09 daily mean temperature (5-km grids) derived from monthly averages. Two different methods to disaggregate monthly temperature into daily data were tested:
  (i) Constant during the month (UKCP09-temp monthly I).
  (ii) Interpolated using pchip method (UKCP09-temp monthly II).'

For illustrative purposes, the following figure will also be added to the supplementary information:

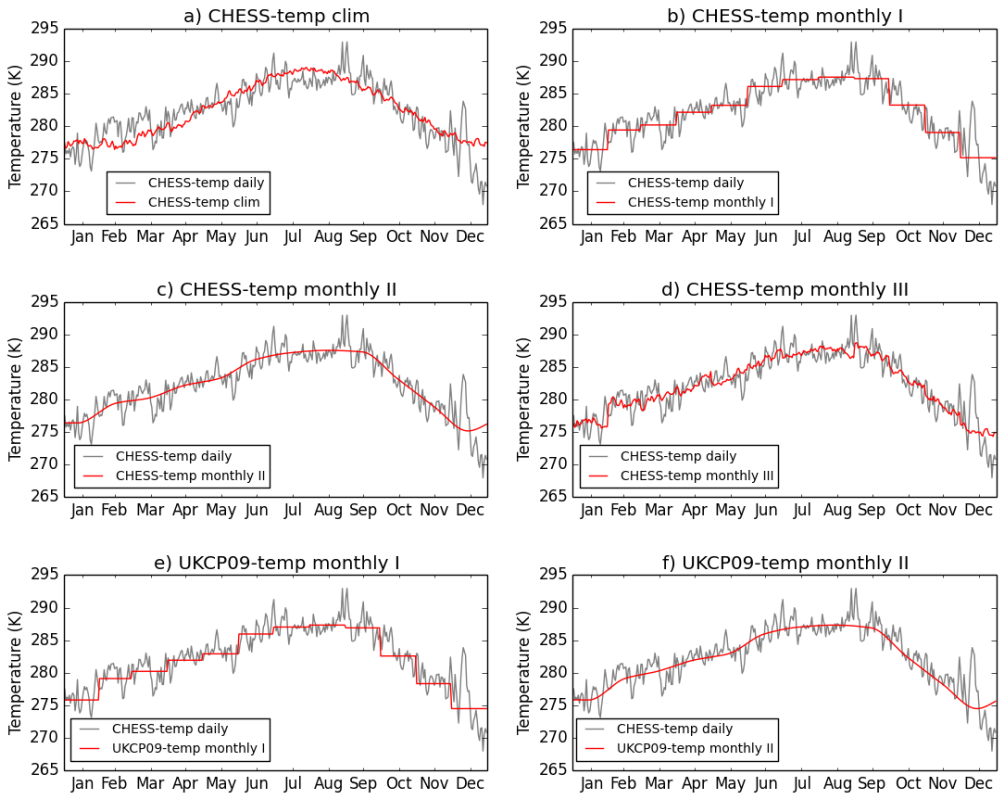

Fig.XX: Illustration of the different input temperature data used. On all plots, the grey line is the daily CHESS temperature data (CHESS-temp daily). The alternative temperature input data tested in the current study are: a) CHESS-temp clim, which is the long term average (1961-1990) of daily mean temperature, derived from CHESS-temp daily; b) CHESS-temp monthly I, which is monthly temperature (from CHESS) disaggregated to daily uniformly for each month; c) CHESS-temp monthly II, which is monthly temperature (from CHESS) disaggregated to daily using pchip interpolation method; d) CHESS-temp monthly III, which is monthly temperature (from CHESS) disaggregated to daily using CHESS daily mean temperature climatology pattern. Effectively, a) and d) are different because in d), the time series have been shifted for each month so that the monthly mean temperature matches the observed monthly temperature. Finally, e) and f) are respectively the same as b) and c), but using UKCP09 monthly temperature instead of CHESS.

Line 29 – are 7 daily datasets derived?

CHESS-temp daily is an existing dataset (https://catalogue.ceh.ac.uk/documents/b745e7b1-626c-4ccc-ac27-56582e77b900). The other 6 daily datasets are manipulated version of existing datasets (CHESS or UKCP09).

Page 5 – Clearer signposting needed in introducing methods. Please link to section in which the detail can be found.

Clearer signposting will be added:

To produce the PET gridded reconstruction product, first a set of seven temperature-based PET equations (presented in section 3.1) were evaluated. These were tested using four different calibration strategies (section 3.2) (in addition to the non-calibrated equations), and seven different temperature input datasets (section 2.1). Once the best combination of equation/calibration strategy/temperature input data was selected (section 3.3.1), the actual gridded PET reconstruction was produced. Test assessing the performance were then carried out on the final gridded product (section 3.3.2), using a range of performance metrics to quantify the reliability of the product in different locations. Figure 1 summarises the different steps of the work.

Line 13 – am not sure about the use of quality assessment tests here – to me these mean homogeneity tests which is not the case in this paper. I think you are assessing performance?

'Quality assessment tests' will be replaced by 'tests assessing the performance' to avoid confusion.

Lines 17-20 – you can delete the first two sentences – repetition Rather than commencing with 'Four main temperature-based equations were evaluated. . .' Start with seven and then differentiate.

OK. Will be changed in revised manuscript

Line 23 – what do you mean by a calibration procedure?

We just mean that the parameters can be calibrated to represent local conditions. To avoid confusion, we will rephrase to: "Each contains a number of parameters representative of the climatic region where the equation was originally developed, which can be calibrated to match the climatic regime of the UK."

Line 28 – min max temps not always available historically either.

We selected low-data demanding methods that could be easily reproduced and extended in cases of minimal data availability. We also considered the applications of this method for forecasting. For the UK, the Met Office currently produces average UK temperature forecasts (for 1-3months) which are used for the production of the UK Hydrological Outlook. Minimum and maximum temperatures are not included in the current seasonal forecasts, which is why we limited our study to temperature-based PET equations that uses mean temperature as input.

Page 6 – line 2 – what do you mean by time efficient?

Assessing the results at catchment scale is computationally much quicker than at individual pixel-scale. This is what we meant by time efficient, but we will remove in revised manuscript to avoid confusion, as it is not essential to the context.

Line 4 – can you provide some indication of range of catchments – size etc.

We will add the following table as a csv file in the supplementary information to provide some information on the catchments (a full version of the table will be added in the final version). We will also add this sentence to the text:

"Table XX in the supplementary information shows the catchments with some of their catchment characteristics."

Table XX: Summary statistics of six catchment characteristics for 306 study catchments. Area, Median elevation, and Base Flow Index (BFI) were retrieved from the UK NRFA. Mean annual Q, P, and PET (based on CHESS-PE) were taken from Harrigan et al. (2017).

Station number	Station name	Easting	Northing	Area (km2)	Median elevation (masl)	BFI (-)	Mean Q (mm yr-1)	Mean P (mm yr-1)	Mean PET (mm yr-1)	Calibration catchment
3003	Oykel at Easter Turnaig	240300	900100	330.7	272.9	0.22	1536.887142	1985.85	403.13	NO
6007	Ness at Ness-side	264500	842700	1839.1	348	0.6	1603.942109	1970.86	394.73	NO
6008	Enrick at Mill of Tore	245000	830000	105.9	335.7	0.3	989.9773598	1433.67	400.15	NO
7001	Findhorn at Shenachie	282600	833500	415.6	558.6	0.36	1124.794091	1288.19	392.24	NO
7002	Findhorn at Forres	301900	858400	781.9	407.9	0.39	836.9534829	1130.40	400.38	YES
7003	Lossie at Sheriffmills	319400	862600	216	196	0.54	413.871875	897.91	427.85	NO
7004	Nairn at Firhall	288200	855100	313	258.9	0.45	557.417562	1032.34	405.68	NO
7005	Divie at Dunphail	300500	848000	165	312.5	0.41	547.0210145	918.76	410.11	NO
8004	Avon at Delnashaugh	318500	835200	542.8	491.6	0.55	862.4872887	1145.22	391.53	NO
8005	Spey at Boat of Garten	294700	819200	1267.8	496.5	0.59	763.3275078	1398.80	386.41	NO
8006	Spey at Boat o Brig	331900	851800	2861.2	419.9	0.6	744.3862725	1198.95	393.98	NO

Reference: Harrigan, S., Prudhomme, C., Parry, S., Smith, K., and Tanguy, M.: Benchmarking Ensemble Streamflow Prediction skill in the UK, Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2017-449, in review, 2017.

Line 13 – sentence beginning this line is long and confusing – break it into two.

The sentence will be split:

"Therefore, four calibration strategies, which are graphically represented in Fig. 3, were considered. The simplest one consists in a global parameterisation leading to a single equation for all 43 catchments. In the most complex approach, a local and monthly parametrisation leads to 12 equations for each of the 43 catchments."

Line 20 – delete of similar length

Ok, will be deleted

Line 21 – were the assumptions of OLS checked?

Yes, they were. A sentence will be added in the main manuscript summarising the results, and a more detailed version will be added to the supplementary information.

The assumption of homoscedasticity (variance of residual is constant) is violated, as the Breusch Pagan test rejects the null hypothesis. The residuals (or errors) are larger when PET is large (spring/summer) than when PET is small (autumn/winter), as is hinted from Fig. A1 of the supplementary information. This is inevitable, as we know that the full variability of PET can't be explained solely by temperature, and other climatic variables, which have seasonal variations, such as wind speed, cloud cover, humidity, etc., aren't taken into account in temperature-based equations, and have a strong influence in PET.

This is why we have tried the 12P-ind and 12P-GB calibration approach (Fig. 3), which removed the heteroscedasticity problem by calibrating the equations separately for each individual month (Breusch Pagan test does not reject the null hypothesis in this case, suggesting homoscedasticity assumption is correct). However, the resulting calibrated equation proved not to be any superior to the globally calibrated version which violated the homoscedasticity assumption, according to our two performance assessment metrics NSE and MAPE (as shown in Fig. 4).

Heteroscedasticity does not affect parameter estimates (as variance, although not constant, is unbiased in our case), but it does affect the estimate of standard errors and confidence intervals. However, we are not using these, and rather we assess the output using a range of other performance metrics (detailed in section 3.3).

In summary, although the assumption of homoscedasticity is being violated, we've seen that applying a monthly calibration (through 12P-ind and 12P-GB approaches), which removes the heteroscedasticity, does not improve results. Moreover, the performance of results is fully assessed using other metrics, therefore we believe that the violation of this assumption does not pose any issue in the application of OLS in our particular case.

In addition, because the variance of the residuals is not constant, we provide monthly MAPE values together with the datasets so that the users are aware of the uncertainty which varies according to the season.

The Durbin-Watson (DW) statistics value of 1.3 also suggests a certain degree of autocorrelation in the residuals, which is common when working with time series. But again, as with heteroscedasticity, autocorrelation tends to underestimate the standard errors, but do not bias the OLS coefficient estimates. As the standard errors and interval of confidence are not used, the moderate degree of autocorrelation does not pose problem in our particular case.

Line 28 – can you call this forcing data – suggest temperature data

OK, will be changed

Page 7 line 7 – this sentence needs reworking- what about hydrological models? Is NSE a concept? Why might NSE be suited to assessing PET?

The Nash Sutcliffe Efficiency (NSE) coefficient was initially developed to assess hydrological models (Nash and Sutcliffe, 1970), but has since then also been used widely to evaluate PET models (Spies et al., 2015, Ershadi et al., 2014, Srivastava et al., 2013, Guerschman et al., 2009, Schneider et al., 2007, Liu et al., 2005). NSE, which is also referred to as Mean Square Error skill score (MSESS) in the forecasting community, looks at how much superior a given model is in predicting a variable (here: PET) compared to the long term average (climatology).

In the manuscript, we will make sure the concept is better explained, and we will add some of these references to illustrate the wide use of NSE for PET model evaluation in the literature.

References:

Spies, R.R., K.J. Franz, T.S. Hogue, and A.L. Bowman, 2015: Distributed Hydrologic Modeling Using Satellite-Derived Potential Evapotranspiration. J. Hydrometeor., 16, 129-146, https://doi.org/10.1175/JHM-D-14-0047.1

A. Ershadi, M.F. McCabe, J.P. Evans, N.W. Chaney, E.F. Wood, 2014, Multi-site evaluation of terrestrial evaporation models using FLUXNET data, Agricultural and Forest Meteorology, Volume 187, Pages 46-61, ISSN 0168-1923, https://doi.org/10.1016/j.agrformet.2013.11.008.

Srivastava, P. K., Han, D. , Rico Ramirez, M. A. and Islam, T. (2013), Comparative assessment of evapotranspiration derived from NCEP and ECMWF global datasets through Weather Research and Forecasting model. Atmos. Sci. Lett., 14: 118-125. doi:10.1002/asl2.427

Juan Pablo Guerschman, Albert I.J.M. Van Dijk, Guillaume Mattersdorf, Jason Beringer, Lindsay B. Hutley, Ray Leuning, Robert C. Pipunic, Brad S. Sherman, 2009, Scaling of potential evapotranspiration with MODIS data reproduces flux observations and catchment water balance observations across Australia, Journal of Hydrology, Volume 369, Issues 1–2, Pages 107-119, ISSN 0022-1694, https://doi.org/10.1016/j.jhydrol.2009.02.013.

Schneider, K., Ketzer, B., Breuer, L., Vaché, K. B., Bernhofer, C., and Frede, H.-G.: Evaluation of evapotranspiration methods for model validation in a semi-arid watershed in northern China, Adv. Geosci., 11, 37-42, https://doi.org/10.5194/adgeo-11-37-2007, 2007.

Siqing Liu, Wendy D. Graham, Jennifer M. Jacobs, 2005,Daily potential evapotranspiration and diurnal climate forcings: influence on the numerical modelling of soil water dynamics and evapotranspiration, Journal of Hydrology, Volume 309, Issues 1–4, Pages 39-52, ISSN 0022-1694, https://doi.org/10.1016/j.jhydrol.2004.11.009.

Page 8 – why these assessment criteria – later it becomes clear but state here.

We will add some text to explain the choice of metrics:

"These six metrics were chosen as they assess different aspects of the modelled data. NSE looks as how much better our model is in predicting PET compared to the long term average (climatology), MAPE gives an indication of the uncertainty, r informs about how well the modelled PET fits the observed values (or 'proxy' to observed in our case), β tells us whether the estimations are biased, VR whether the spread of the estimated values matches the observed spread, and finally KGE informs on the combined effect of r, β and VR."

Line 24 – 1P-GB introduced first time – at least link to the figure.

Will be linked to figure 3

Page 9 – line 5-8 sentence too long, too many commas.

Will be rephrased to:

"A surprising result is that, in the absence of any climate data available, calibrating McGuinness-Bordne equation with CHESS-temp clim (long-term daily temperature climatology) outperforms using CHESS-PM climatology."

Line 12 delete in conclusion – this is not the conclusion

Will be deleted

Am left wondering if reduction in temperature station density back in time is evident and if this affects results.

As detailed in our answer to the first specific comment, the change in station density should only have a minor effect.

Page 10 – line 16 no need to present this correlation coefficient

OK, will be removed.

Line 31 – give us the values of the more moderate performance

Ok will be added: For daily values, the performance is more moderate (NSE > 0.4, r > 0.8 and KGE > 0.7).

Page 11 – discussion and limitations needs to include fuller assessment of uncertainties.

The discussion around the uncertainties coming from the underlying temperature data and change in network density will be added.

Line – 12 – please state what PM is here

PM is defined earlier (page 5, line 3), but we will remind the reader what PM is here again (Penman Monteith).

Line 18 – replace would most likely be with are

Ok, will be changed

Line 19 Great Britain? This include NI or Scotland?

Great Britain included Wales, England and Scotland (but excludes Northern Ireland). United Kingdom (UK) includes Northern Ireland in addition to all the previous ones. We will make sure this is explained clearly in the revised manuscript for non-British readers.

Line 19 – when and where such high resolution...

OK, will be changed

Line 23 – replace unique with calibrated

OK, will be changed

Line 26 – e.g. provide guidance

Checking the performance metrics in the area of interest is highly encouraged in order to assess the reliability of the daily PET estimation.

Page 12 – line 11 – move this finding (vi) to higher prominence.

Ok, will be moved up in the list.

Line 14 – replace metrics grid with gridded metrics

Ok, will be changed.

Line 19 – The name of the dataset is a little misleading potentially – as it is stated it reads that it is calibrated for the UK over the period 1890-2015. Consider rewording.

The title is linked to the DOI, so can't be changed. However, the metadata record and supporting documentation (which can be found here: https://catalogue.ceh.ac.uk/documents/17b9c4f7-1c30-4b6f-b2fe-f7780159939c) has been amended so that it is clear that the equation was calibrated for the period 1961-1990.

I have not checked the references

We have checked the references, but we will make sure we will double-check them in the final version.

Captions Fig 1 – caption needs to be more informative and help the reader interpret this complex figure

A new, more informative caption has been added to the new Figure 1 (page 4 of this document)

Fig 2 – same, could be more informative

New caption: Figure 2: Maps of the boundaries and outlets of a) catchments that were used to calibrate the PET equations and to calculate the performance metrics of the PET equations (described in section 3.3.1), and b) catchments that were used to carry out the assessment of the final PET grids using the performance metrics described in section 3.3.2.

Fig 3 – same point

New caption for Fig.3: Schematic of the calibration strategies. Four calibration approaches were considered to calibrate the PET equations: from local and monthly parametrisation leading to 12 equations for each of the 43 catchments (12P-ind), to a global parameterisation leading to a single equation for all 43 catchments (1P-GB).

Fig 5 – in caption 4e should be 5e. What is upper VR range, please relate to part of the text where indices are described.

OK, caption will be corrected, and reference to section mentioned (section 3.3. Evaluation).

Fig 6 – 5d should be 6d and same for 5e

Ok caption will be corrected.