Developing high-quality meteorological data for East and West Africa from merged sources

Working Paper No. 45

CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS)

Justin Sheffield
Nathaniel W. Chaney

RESEARCH PROGRAM ON Climate Change, Agriculture and Food Security

Working Paper

Developing high-quality meteorological data for East and West Africa from merged sources

Working Paper No. 45

CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS)

Justin Sheffield
Nathaniel W. Chaney

Correct citation:
Sheffield, J., and N. Chaney, 2013. Developing high-quality meteorological data for East and West Africa from merged sources. CCAFS Working Paper No CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS). Copenhagen, Denmark. Available online at: www.ccafs.cgiar.org

Titles in this Working Paper series aim to disseminate interim climate change, agriculture and food security research and practices and stimulate feedback from the scientific community.

This document is published by the CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS), which is a strategic partnership of the CGIAR and the Earth System Science Partnership (ESSP).

Contact:
CCAFS Coordinating Unit - Faculty of Science, Department of Plant and Environmental Sciences, University of Copenhagen, Rolighedsvej 21, DK-1958 Frederiksberg C, Denmark. Tel: +45 35331046; Email: ccafs@cgiar.org

Table of Contents

Abstract

Assessments of agricultural productivity and food security require process-based crop models to provide predictions of yields and diagnose past variations in the context of anthropogenic and climate factors. These models need detailed meteorological data as input, including precipitation, temperature, humidity, solar radiation and windspeed. This project aimed to apply existing methods to merge in situ, remotely sensed and modeled data sources in East and West Africa to produce high-quality daily meteorological data over at least 30 years. Specific objectives included: evaluation of the error structure of the dataset, its temporal and spatial characteristics and consistency and its suitability for forcing crop models, and to provide a framework for merging new data, in particular from the local stations of regional African partners, ensuring consistency across time and space and among variables, as well as the best use of information.

The work successfully created a 10 kilometre, daily meteorological dataset for East and West Africa for the period 1979–2008, based on the National Centers for Environmental Prediction-National Center for Atmospheric Research (NCEP–NCAR) reanalysis (NNR), merged with observational datasets, including the monthly gridded precipitation and temperature product of the University of East Anglia's Climate Research Unit (CRU), the NASA Langley Surface Radiation Budget (SRB) product, and station data from the Global Summary of the Day (GSOD) database.

Keywords

Daily meteorological dataset; data merging framework; crop modeling; East Africa; West Africa.

About the authors

Both authors are in the Department of Civil and Environmental Engineering at Princeton University, Princeton, NJ, 08540, USA.

Dr. Justin Sheffield is a research hydrologist.

Nathaniel Chaney is a graduate student in the Environmental Engineering and Water Resources Program.

List of abbreviations and acronyms

AVHRR	Advanced Very High Resolution Radiometer
CDF	Cumulative Distribution Function
CM-SAF	Satellite Application Facility on Climate Monitoring
CRU	(UK University of East Anglia's) Climate Research Unit
ETCCDI	Expert Team on Climate Change Detection and Indices
EUMETSAT	European Organisation for the Exploitation of Meteorological Satellites
GERB	Geostationary Earth Radiation Budget
GSOD	Global Summary of the Day
MSG	Meteosat Second Generation
NCAR	National Center for Atmospheric Research
NCDC	National Climatic Data Center
NCEP	National Centers for Environmental Prediction
NLDAS-2	North American Land Data Assimilation System- Phase 2
NNR	NCEP–NCAR reanalysis
NOAA	National Oceanic and Atmospheric Administration
SEVIRI	Spinning Enhanced Visible and Infrared Imager
SRB	Surface Radiation Budget
TIROS	Television Infrared Observation Satellite
TOVS	TIROS Operational Vertical Sounder

1. Introduction

Assessments of agricultural productivity and food security require process-based crop models to provide predictions of yields and diagnose past variations in the context of anthropogenic and climate factors. These models need detailed meteorological data as input (also known as forcing data), including precipitation, temperature, humidity, radiation and windspeed. Until recently, agricultural assessments at large space and time scales have been hampered by the lack of detailed and accurate meteorological data to force the models. Traditionally, necessary meteorological data were only available from local station observations, which are not pervasive across all global areas and certainly not at the spatial and temporal resolutions that are required by crop models for most applications. In developing regions, such as most of Africa, the density of station data is much lower. Coupled with the lack of temporal extent and consistency in the majority of observations, this has made the development of forcing datasets using observations alone unsatisfactory. The accuracy of the meteorological data is also important, as first order errors in modeled states have been demonstrated to be due to inaccurate forcings and especially in precipitation (e.g. Robock et al., 2003; Challinor et al., 2005). The conclusion of these and other studies is that accurate forcings are necessary to provide accurate simulations when compared to observations. For climate change studies, the temporal consistency of the forcing data is also of paramount importance.

With the increasing availability of global remote sensing products, the prospects are more promising, although their generally short time period hampers development of long-term datasets required for climate related studies. In the global context, the use of atmospheric reanalysis products may be the only alternative for providing near surface meteorological forcings at high temporal resolution. In contrast to the lack of terrestrial observations, the relative wealth of observations of the atmosphere and sea surface has allowed the emergence of a number of global, long-term, reanalysis datasets such as the NCEP/NCAR, ERA-40, and NASA-DAO reanalyses. These products are constructed using 'frozen' versions of numerical weather prediction and assimilation systems that ingest a variety of atmospheric and ocean observations to provide long-term, continuous fields of atmospheric (and land surface) variables. These first-generation reanalyses have now been enhanced or superseded by second-generation products (e.g. ERA-interim, NARR, MERRA) that use improved physical models and assimilation schemes, higher resolution and novel data sources such as assimilation of observed precipitation.

The power of reanalyses is their consistent and coherent framework for ingesting in situ and remote sensing data into a time- and space-discretized representation of the global land, oceans and atmosphere, in a way that is essentially impossible to achieve directly from observations. However, the direct use of reanalysis fields to force land models is hindered by inherent biases in the atmospheric model and changes in the observing systems that provide data for assimilation. For example, Sheffield et al. (2004) and Ngo-Duc et al. (2005) showed that systematic biases in reanalysis meteorology filter down into modeled land surface variables. Nevertheless, the results of such studies have shown that there is great potential for using hybrid datasets, which combine reanalysis with observation-based datasets to remove biases. This approach retains the consistency and continuity of the reanalysis but constrains it to the best available observation datasets, which are generally available at coarser resolutions and reduced spatial and temporal extents.

We followed this approach to develop a long-term, high-resolution, and high-quality dataset of daily meteorological forcings for East and West Africa that is suitable for crop modeling and other modeling applications. The work builds on previously constructed long-term datasets of precipitation and other meteorological data (Sheffield et al. 2006) for use in driving hydrologic, ecosystem and agricultural models to explore relationships between climate and terrestrial processes. The dataset is also suitable for analysis of long-term changes in meteorology and some examples are shown for daily extremes of temperature and precipitation.The specific objectives were to:

- implement existing methods for merging in situ, remotely sensed and modeled data to generate a dataset of daily meteorological variables over at least 30 years, suitable for forcing typical crop models;
- evaluate the error structure of the dataset, its temporal and spatial characteristics and consistency and its suitability for forcing crop models;
- provide a framework for merging new data, in particular from the local stations of regional African partners, ensuring consistency across time and space and among variables, as well as the best use of information.

We successfully created a gridded, 10 km daily meteorological dataset for East and West Africa for the period 1979–2008. This is based on the NCEP–NCAR[1] reanalysis (NNR), merged with the monthly gridded precipitation and temperature product of the University of East Anglia's Climate Research Unit (CRU) and the NASA Langley Surface Radiation Budget (SRB) product. The CRU products were evaluated for temporal inconsistencies attributable to changes in contributing gauges; adjustments were made to remove step changes and ensure consistency among variables, such as downward long-wave and humidity/temperature, and between temperature and humidity to maintain relative humidity. Empirical adjustments to downward short- and long-wave surface fluxes were made to achieve consistency with precipitation. Scaling down in space from the coarse resolution of the NNR, CRU and SRB products to 10 km resolution was carried out, accounting for elevation effects.

The dataset was evaluated against observations from the Global Summary of the Day (GSOD) of the US National Climatic Data Center (NCDC) and a range of statistics on timescales from days to years. A method for assimilating station data into the gridded dataset was developed and tested and was used to merge available station data into the full 1979–2008 gridded dataset. Products such as annual, monthly and daily statistics, and extreme values and potential evaporation, were calculated to demonstrate the utility of the dataset.

1 National Centers for Environmental Prediction–National Center for Atmospheric Research.

2. Project findings

The following sections of this Working Paper explain the key findings from the tasks involved in the project.

2.1 Acquire and evaluate contributing datasets

The NNR, CRU, SRB and various evaluation datasets were obtained for their most recent available time period and versions. The NNR product was obtained up to 2010 for near-surface meteorological variables at six-hour intervals. The latest version of the monthly CRU dataset (TS3.1) was obtained for a range of variables, including precipitation, temperature, cloud cover and vapour pressure. The SRB (V3.0) surface-radiation dataset was available for 1983–2007. We also obtained the EUMETSAT CM-SAF geostationary, 3 km daily surface-radiation products. These are based on SEVIRI/GERB instruments on the MSG satellite, and AVHRR sensors on MetOp and National Oceanic and Atmospheric Administration (NOAA) satellites, developed by the German meteorological service.

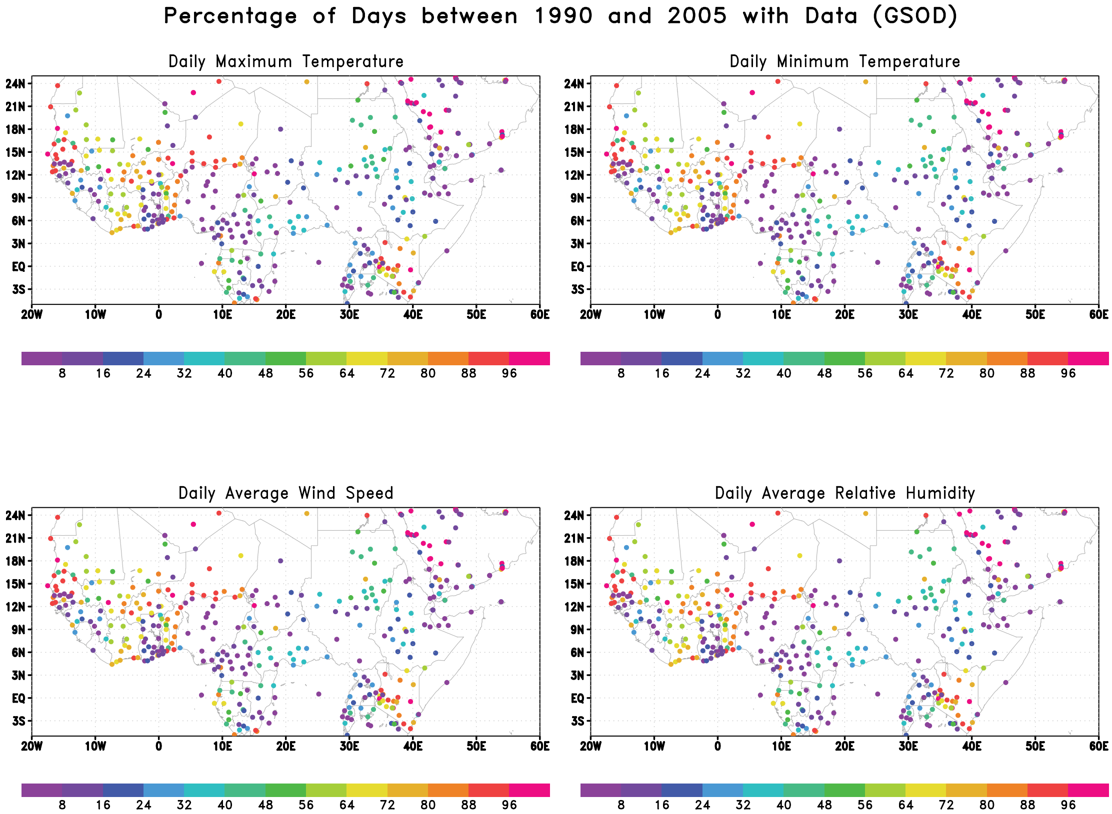
GSOD data were obtained and processed for several hundred stations across East and West Africa. Figure 1 shows the location and number of data records for the GSOD stations as an example for 1990–2005, showing mean, maximum and minimum daily temperatures, humidity and wind speed. The GSOD data were used to evaluate the CRU datasets, which are generally based on much fewer stations than are available, and to demonstrate the assimilation of station data into the merged gridded product.

2.2 Evaluate datasets for robustness and consistency across time and space

The datasets that contribute to the gridded product contain biases because of the low density of contributing stations and changes in stations, satellite sensors and possibly methodology. The CRU monthly temperature data, for example, are gridded from station data that are not uniform in space and time and that may induce step changes and spatial inconsistencies, as well as biases in regions with few stations. The SRB dataset may also contain spurious jumps because of changes in satellite sensors and retrieval algorithms. We tested for these changes (steps and trends in the mean and variability) using regime identification with a moving-window t-test that withstands trends (Rodionov 2004) and applied simple methods to remove any step changes that were found.

Figure 2 shows the average temperature station count in the CRU dataset for 1990–2005 and indicates large regions without station coverage and low temporal coverage where there is a station present. Over the whole period of the CRU dataset (1901–2009), there is considerable variation in the number of available stations with a peak between 1950 and 1990, after which there is drop of about 50%. The CRU data were developed by gridding the available station data based on the length of anomaly correlations, which are up to 1200 kilometres for temperature. The lower panels of figure 2 show the number of stations that contribute to the gridded values and these reflect the location of the stations. For some time periods and regions there are no available station data within the correlation distance and therefore, with no other information, the CRU data are set to 'climatology.'

Figure 1. Percentage of days in the GSOD database 1990-2005



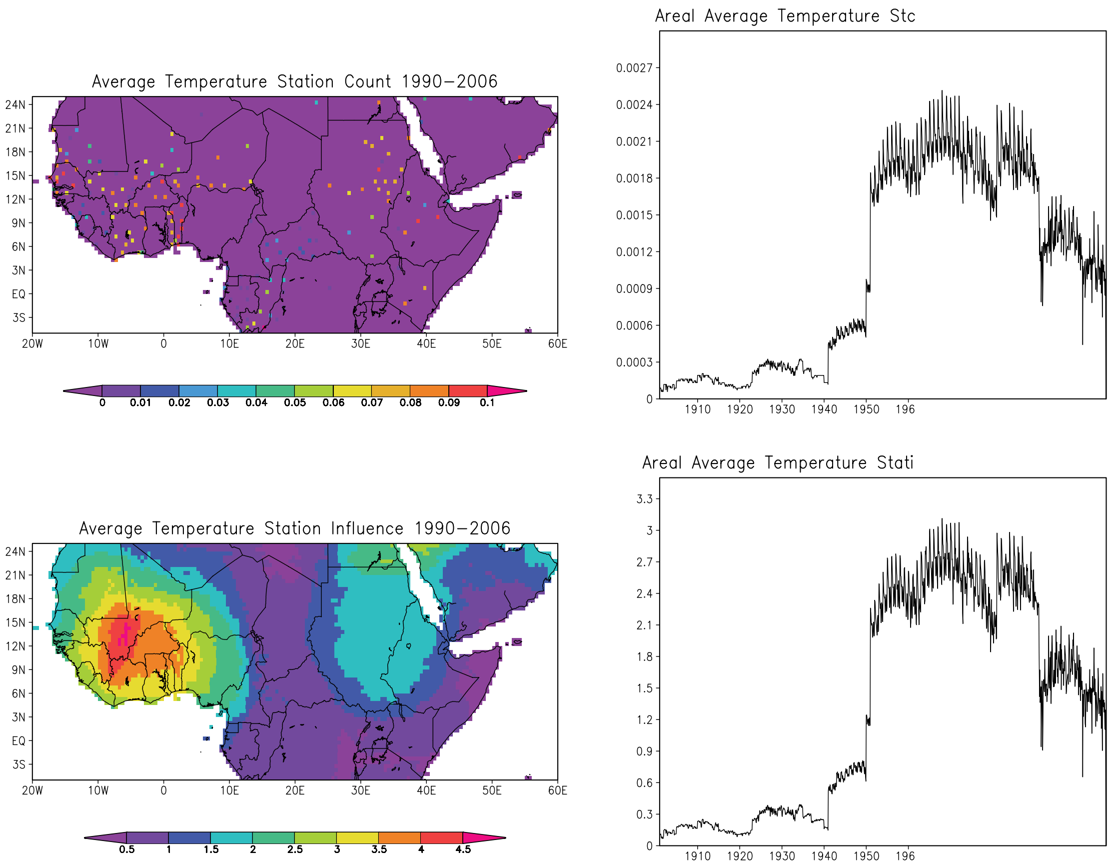
We tested for trends and step changes in the CRU dataset to identify potential problems due to the limited station data. Some examples are given in figure 3 for monthly temperature, which shows the original CRU data anomalies, the number of contributing stations and a comparison with an alternative gridded temperature dataset of Matsuura and Willmott (2010). The figure shows periods with step changes based on the moving-window t-test.

For the first grid cell, the identified step change is attributable to a drop in the number of contributing stations not seen in the alternative dataset. For the second grid cell, the step change does not appear to be associated with a change in stations but is spurious when compared to the alternative gridded dataset. An overview of identified step changes for all variables is given for the gridded product in the next section. Spurious step changes are removed by shifting the data after the step change so that its mean matches the mean of the data before the step.

2.3 Implement existing methods to develop a regional test dataset as proof of concept

The methods of Sheffield et al. (2006) were used to develop an initial gridded dataset for the region at 10 kilometres, daily resolution for 1990–2005. In summary, the NNR was merged with gridded station data (CRU) and satellite products (SRB) at monthly time-step and down-scaled down in space to 10 kilometres, considering elevation effects. Figure 4 shows example fields of the eight variables (mean temperature, maximum temperature, minimum temperature, surface downward solar radiation, surface downward long-wave radiation,

Figure 2. Average number of stations (top) (number of stations per month) and contributing stations (top) (number of stations per month that contribute to the gridded value) for the CRU gridded monthly analysis shown (left) averaged over 1990-2005 and (right) the temporal evolution averaged over the whole region



surface pressure, 2-metre humidity and 10-metre wind speed). This is prior to assimilating GSOD station observations, discussed later.

The gridded data were tested for trends and step changes. Figure 5 shows the results in terms of the maximum step-change index value, where warmer colours indicate changes with higher statistical significance. Some regions, such as central East Africa stand out for shifts in temperature illustrated previously in figure 3. Surface downward long-wave radiation shows a large-scale shift that peaks around mid-1998, which is consistent with the change in the retrieval algorithm for the NOAA's TIROS Operational Vertical Sounder (TOVS), which introduced a drop in atmospheric humidity and temperature. For other variables such as pressure, which for the monthly variability is derived solely from the NNR, the shifts are more spatially coherent and widespread, and therefore likely to be real shifts in the climate regime.

However, relative humidity is also derived from the NRR and shows similar shifts around mid-1998, if more localized. Examination of time series at individual grid cells indicates that these are consistent shifts in the mean, which are simply corrected by matching the mean to that before the shift.

Figure 3. Identification and attribution of shifts in the CRU data for two example grid cells (left and right columns). Potential spurious shifts in the CRU monthly temperature data identified using a 20 month moving-window t-test (top). Number of contributing gauges (middle). Comparison of CRU anomalies with those from the Willmott-Matsuura product (bottom)

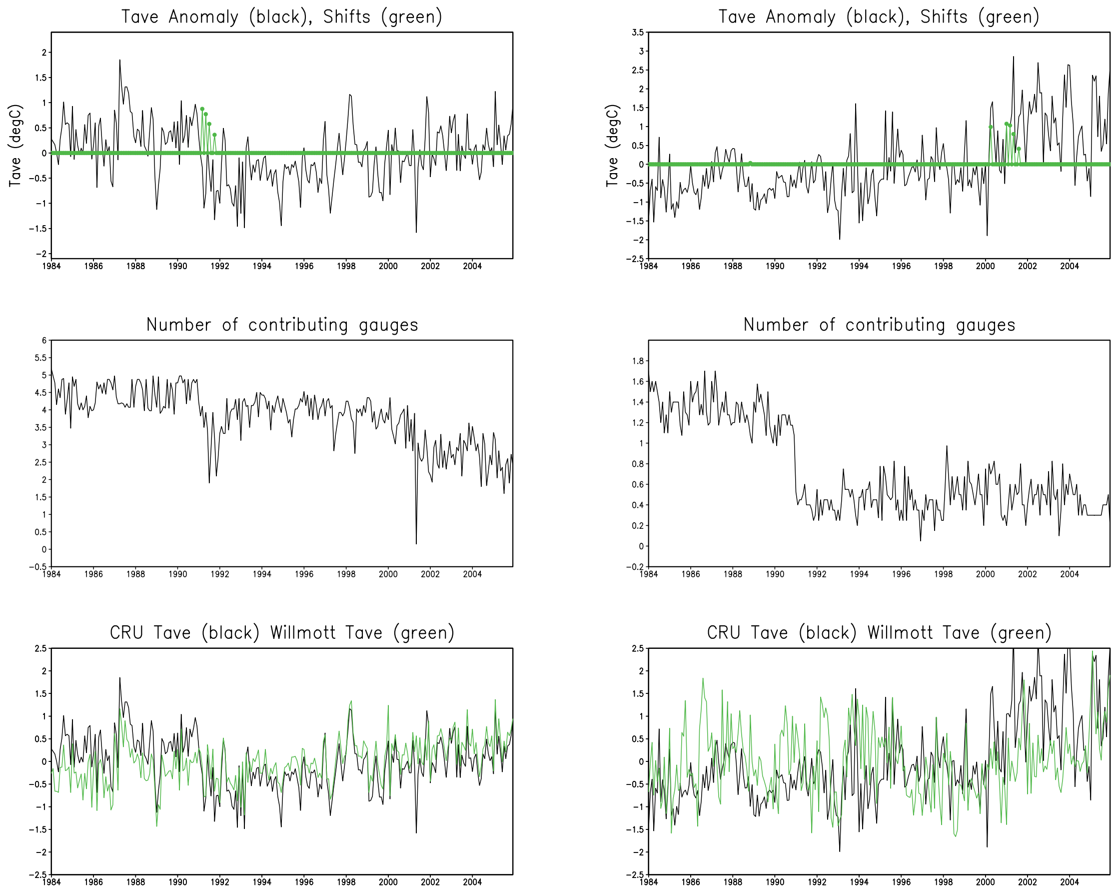


Figure 4. Daily example of 10 km resolution maps of the 8 variables in the demonstration gridded dataset

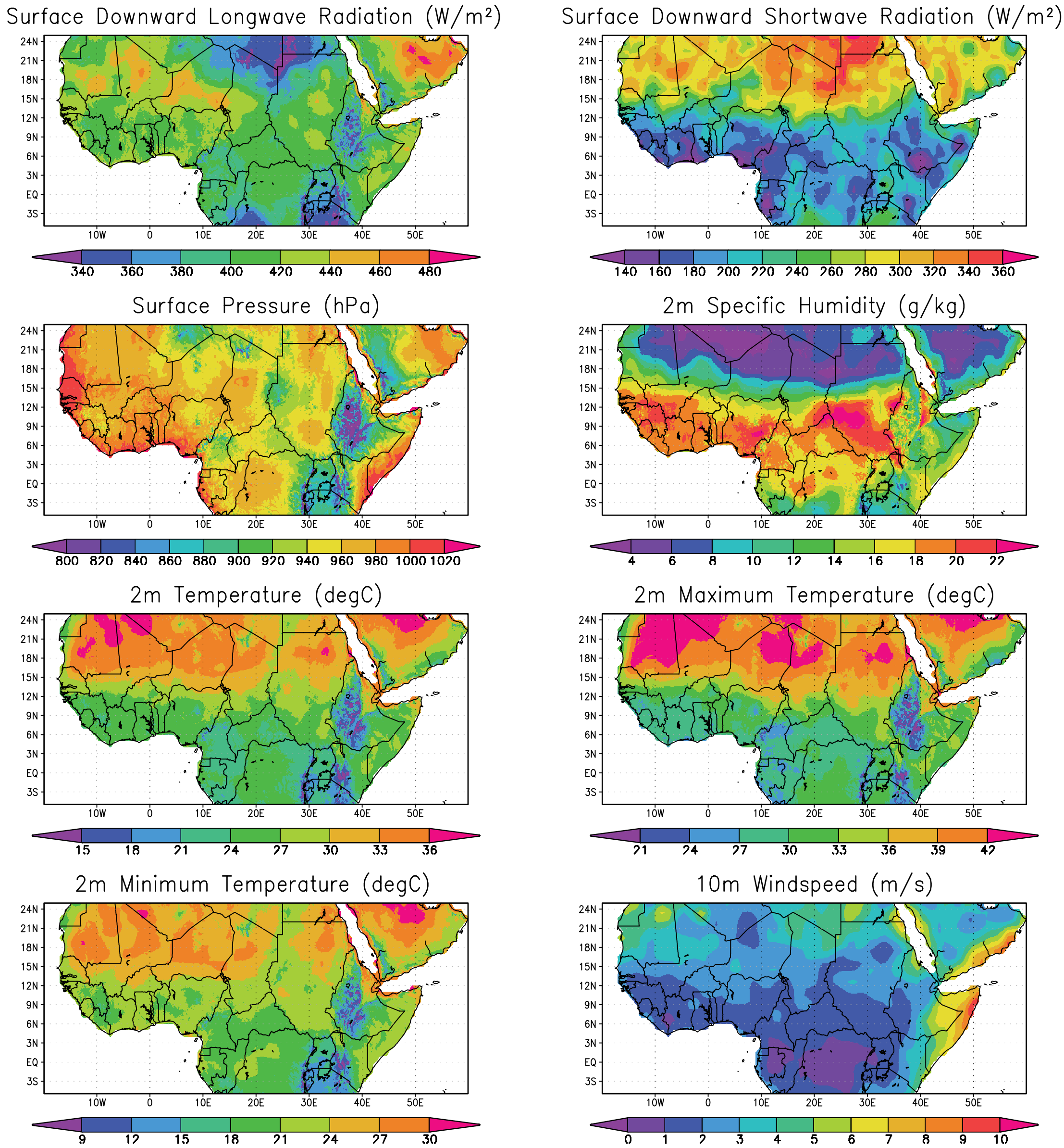
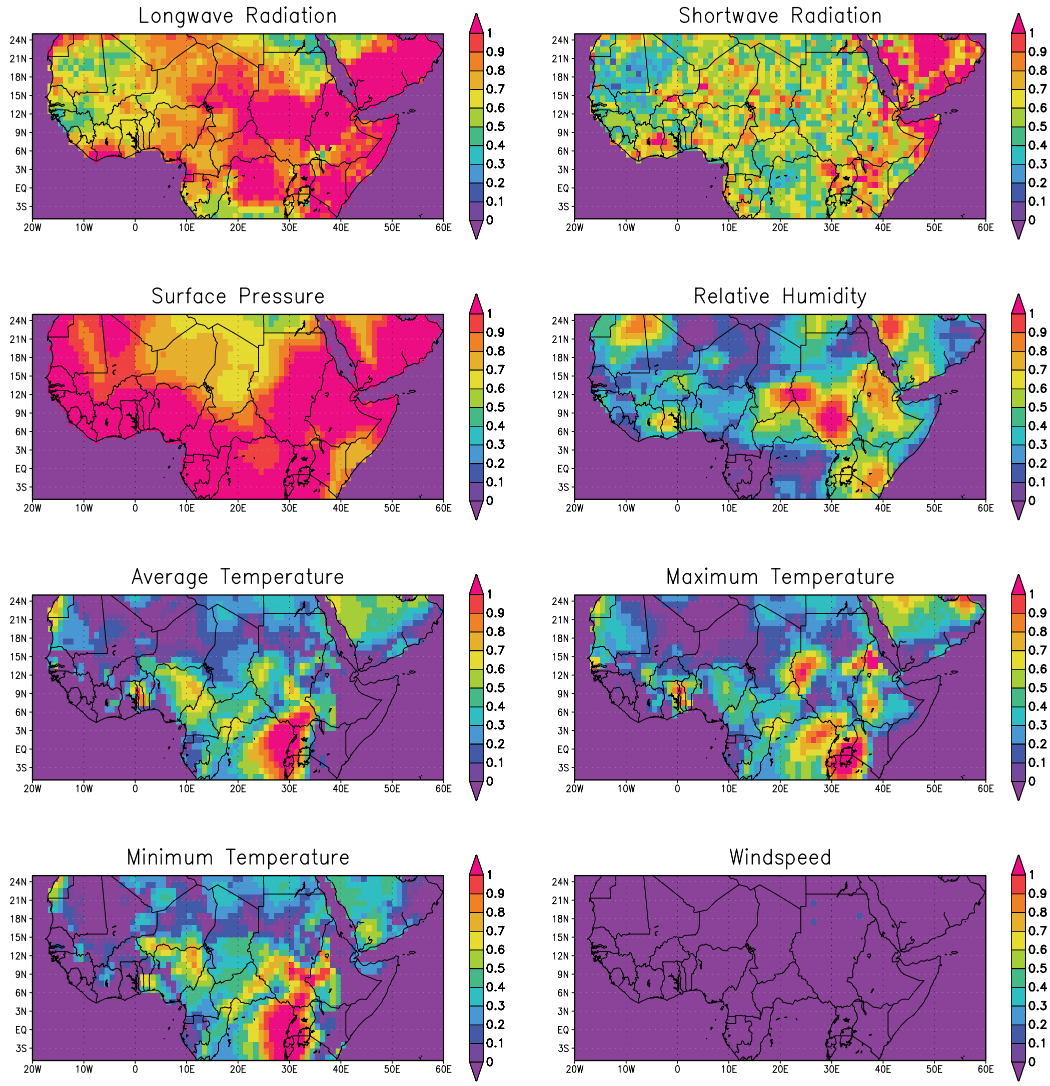


Figure 5. Maximum step change index for the gridded dataset at monthly time step



2.4 Evaluate the data against station observations

The gridded daily dataset was evaluated against the GSOD database, which consists of station measurements from several global and regional databases. The database has reasonable spatial coverage except in the centre (Republic of the Congo, southern Sudan) and far eastern parts of the region (Somalia). Correlations of the gridded dataset with the GSOD on a daily time scale (figures 6 and 7) are reasonably high north of 5°N, but are very low in most of central Africa. Correlations are highest for minimum temperatures and relative humidity and lowest for wind speed, as expected given the high spatial and temporal variability of wind speed. Correlations are higher for monthly maximum and minimum temperatures and humidity, but do not improve much for monthly wind speed.

Absolute errors are shown in figure 8. Daily absolute errors are of the order of 1–2°C across West Africa and 2–3°C across East Africa for maximum and minimum temperature, with maximum errors of about 4°C. Daily absolute errors in wind speed are about 1 m/s across the region with errors exceeding 2 m/s at some stations. For relative humidity, daily errors reach 15% along the coast of Guinea and across the Horn of Africa. Errors generally reduce

Figure 6. Pearson correlation between high-resolution gridded data and GSOD station data for (top) daily and (bottom) monthly time step

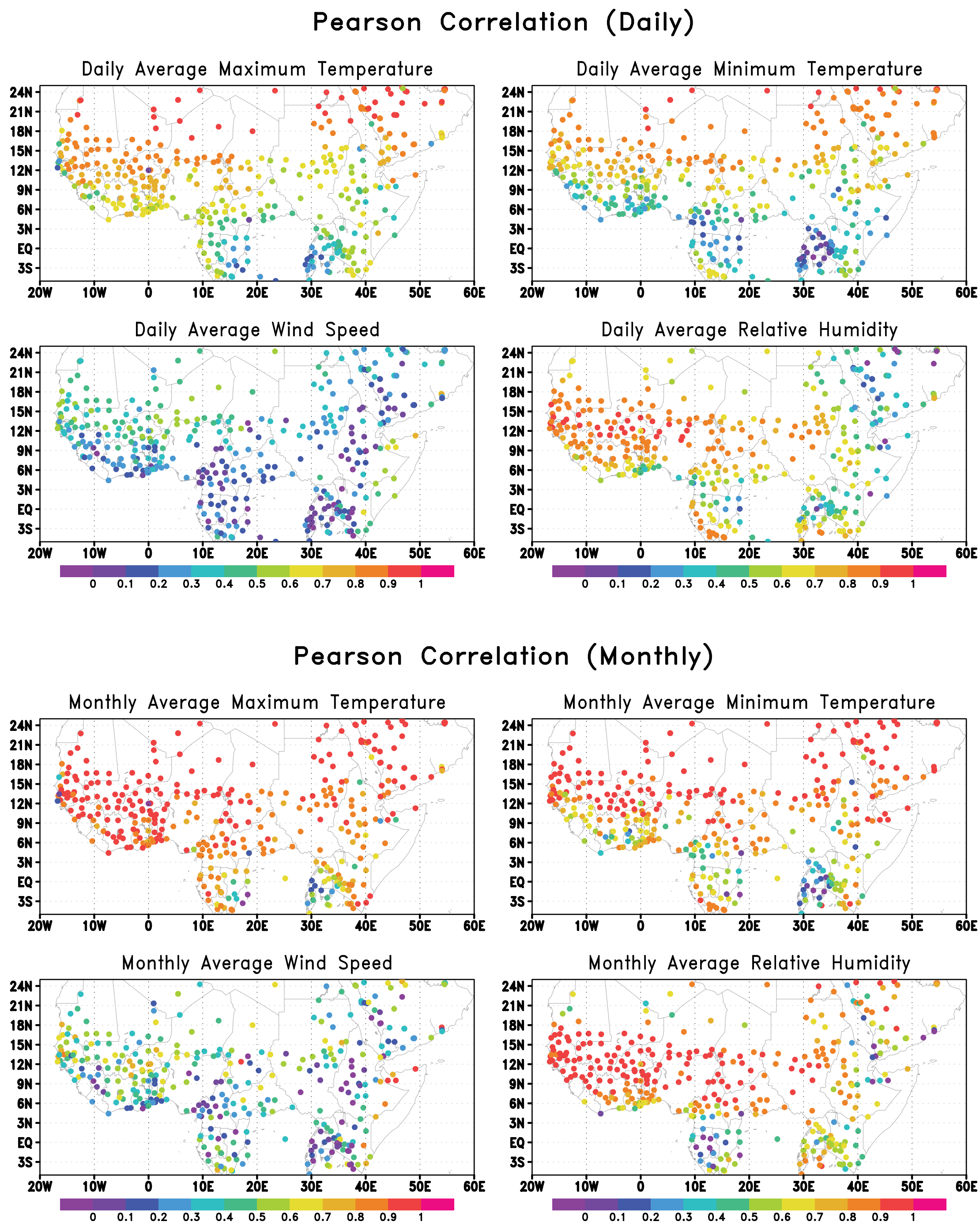


Figure 7. Histograms of Pearson correlation coefficients between the high-resolution gridded data and the GSOD station database for (top) daily data and (bottom) monthly data

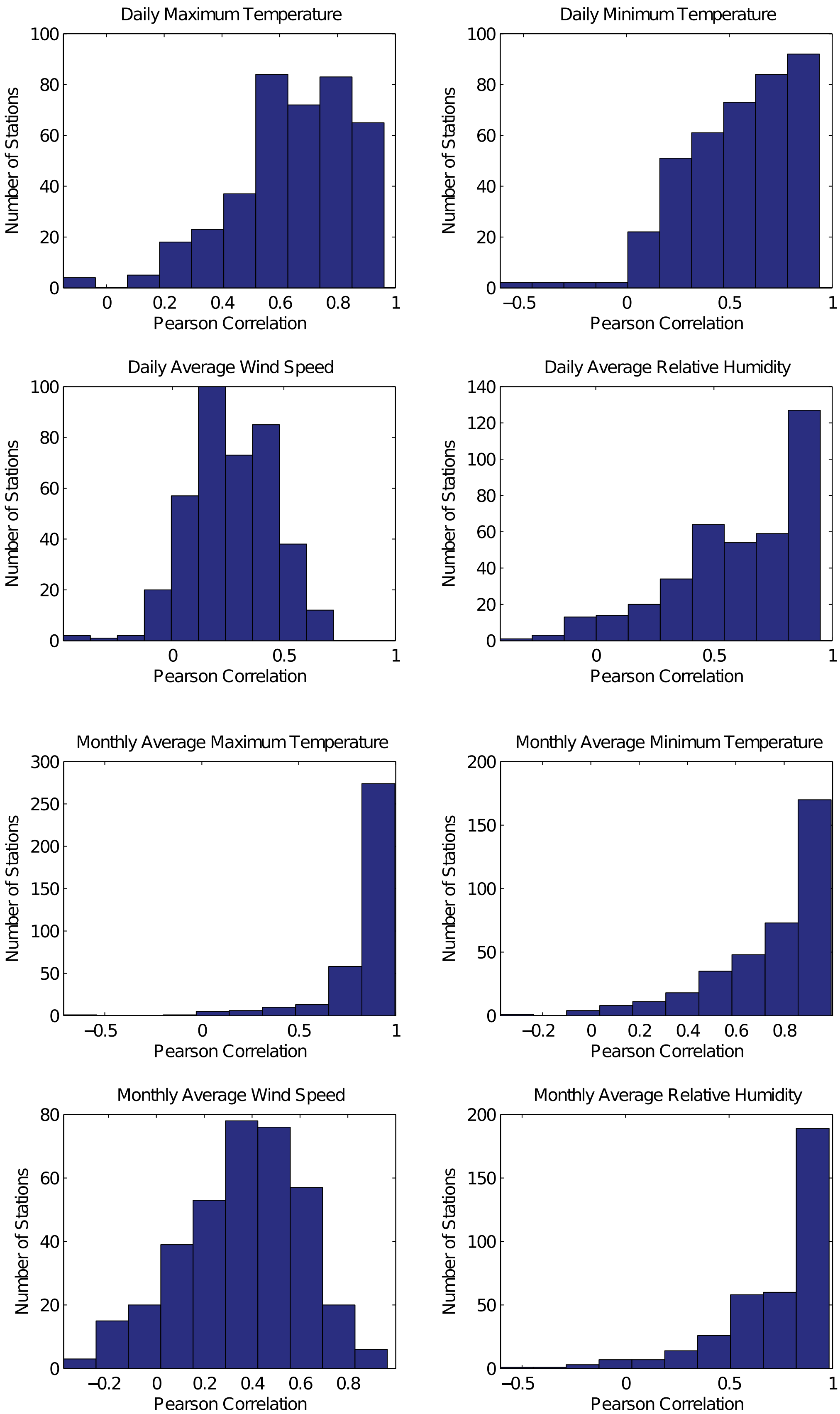
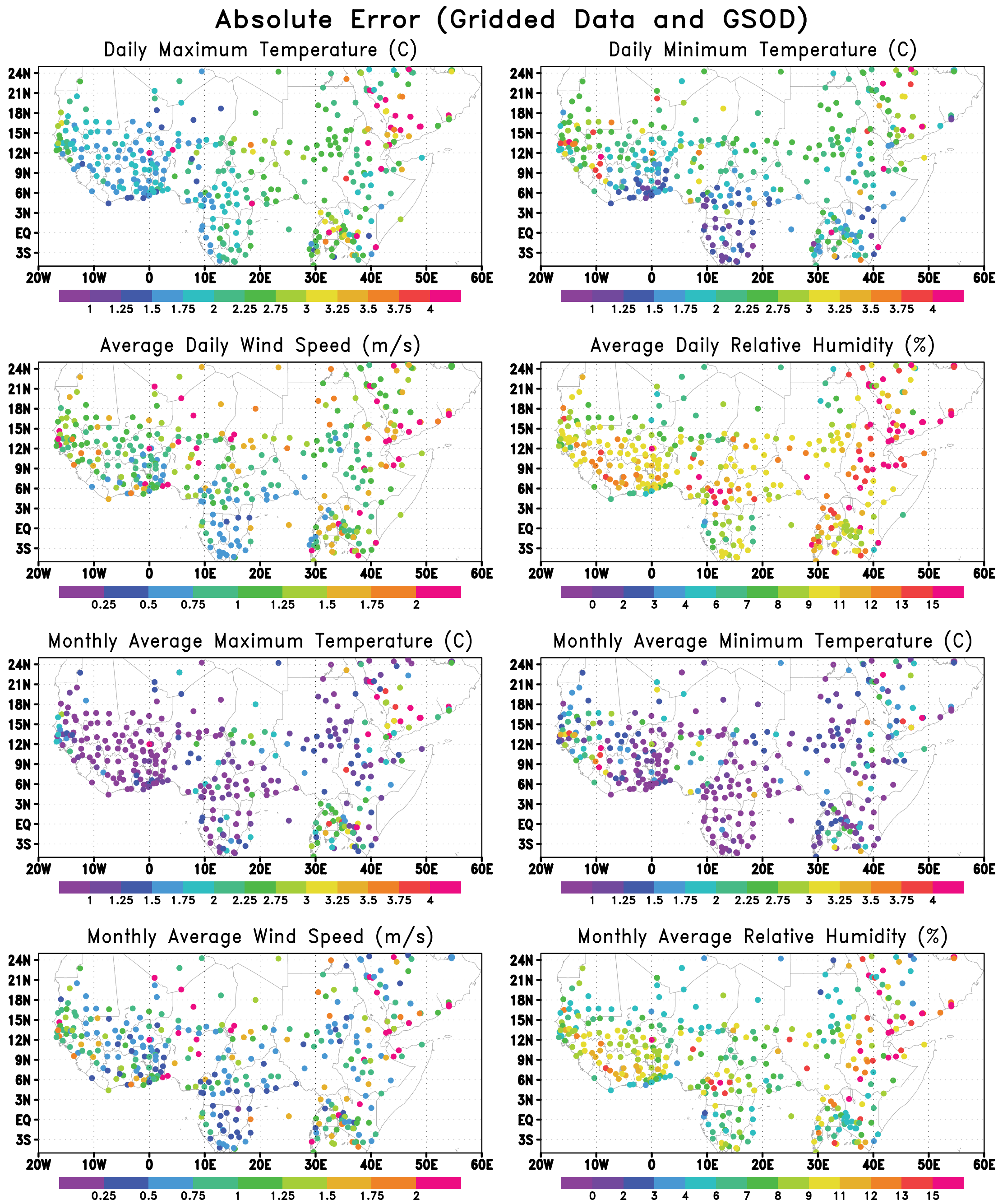


Figure 8. Absolute errors between the gridded dataset and the GSOD station database at daily (top) and monthly (bottom) time scales 1990-2005



on a monthly time scale to at most 1.5°C for temperature, less than 1 m/s for wind speed, and about 10% for relative humidity.

Figure 9 shows error statistics for extreme days, defined as the 10^{th} or 90^{th} percentile in the GSOD dataset. The gridded dataset has too many 'very hot' days (defined as days warmer than the 90^{th} percentile) in the east and a tendency for too few in the west. For minimum temperatures, the gridded data is generally too warm (days warmer than the 10^{th} percentile) and so has too few days with low minimum temperatures. For relative humidity, the gridded data has too many low-humidity days across the region.

2.5 Develop methods for merging station data

A methodology for assimilating station data into the gridded product was developed based on Chirlin and Wood (1982). The errors in the 10 km gridded data were corrected using station values by computing a set of weights based on the spatial relationship between the stations that best combines the correction factors from these stations. These weights were then applied to give a corrected grid value.

In the terminology of data assimilation, the original gridded data are denoted as the 'background field.' The corrected grid-point data value y^* is computed as follows:

$$y^* = y + G(y^d - Hy)$$

Where:

y = background data value
G = gain matrix (weights)
y^d = value of station data
H = measurement matrix
$(y^d - Hy)$ = correction factors

The solution of the gain matrix G is obtained by minimizing the mean squared error of the estimated value y^*. This reduces the problem to the solution of a system of N linear equations, where N is the number of stations.

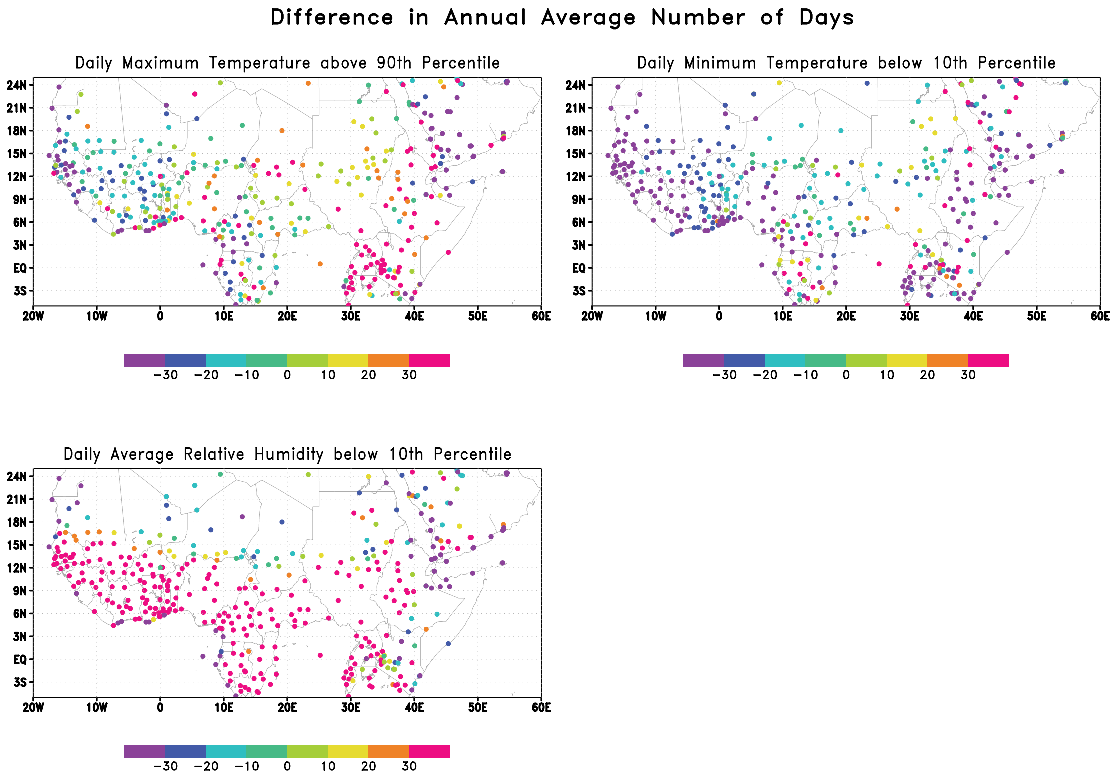
$$G = (P\ HT)\ (H\ P\ HT)^{-1}$$

P is the covariance matrix, which defines the spatial relationship between all the data points of interest: the stations and the grid point to be corrected. It should be noted that given this definition of G, the station errors are negligible and thus all error is attributed to the gridded data set.

In this implementation, P was calculated from the experimental variogram derived from the gridded data set at each time step; stationarity and isotropy were assumed. The experimental variogram defines the relationship between two points as a function of the distance between the points, as follows.

h = distance between two points, C(h) = number of points that are within a certain distance range or 'bin.' In essence, this is the average squared difference between all pairs of points for which the distance between the two points is under a given threshold. A variogram model was fitted to the experimental variogram to ensure a solution to the system of equations defined by G (figure 10).

Figure 9. Difference in the annual average number of days that both the gridded data and GSOD stations surpass a threshold set from the GSOD stations



Assuming spatial stationarity, the covariance function C(h) is computed from the variogram by the following equation:

$$C(h) = \gamma(0) - \gamma(h)$$

The covariance matrix P was then populated by realizations of C(h) from the distance between the points of interest (stations and grid point to be corrected) in the domain. The problem was then reduced to solving for G, and applying these weights to the correction factors to give the optimal solution for each grid cell (y*).

2.5.1 Testing using Oklahoma Mesonet

The method was initially tested over the Oklahoma Mesonet monitoring network in the US, which has a dense distribution of stations and represents one of the most densely monitored regions of the world (figure 11). As such, it robustly tests how the method might perform over Africa where station density is much lower. We tested the merging method for daily temperature extremes and wind speed using the NLDAS-2 gridded meteorological dataset (Xia et al. 2011), which merges reanalysis with gridded observational data for the continental US at 1/8th degree spatial resolution, and is thus similar to our gridded product. Figure 11 shows the influence of increasing station density on the corrected gridded field. It also shows the impact of the derived variogram in restricting the impact of isolated stations (e.g. n stations = 8) to the local area.

Figure 10. Experimental variogram obtained from the NLDAS (see below) daily maximum temperature data on 1 January 2002. A spherical model variogram is then fitted to the experimental variogram to define the spatial relationship.

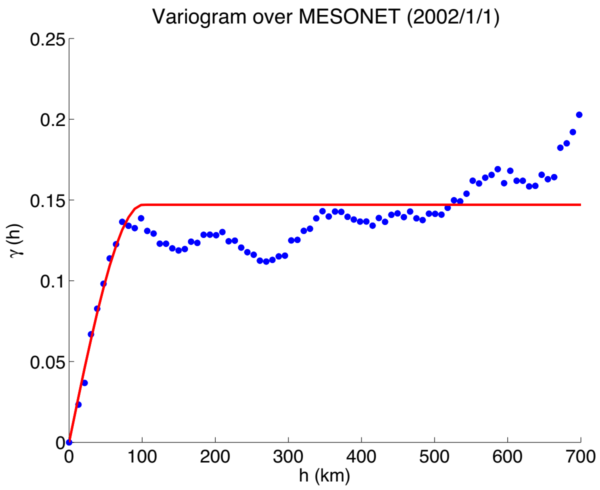


Figure 11. Test of the merging method over the OK MESONET network for 1 January 2002. This figure illustrates the impact of the derived variogram on defining the impact of an individual station on the field. It also shows how increasing the network density increases the accuracy of our product.

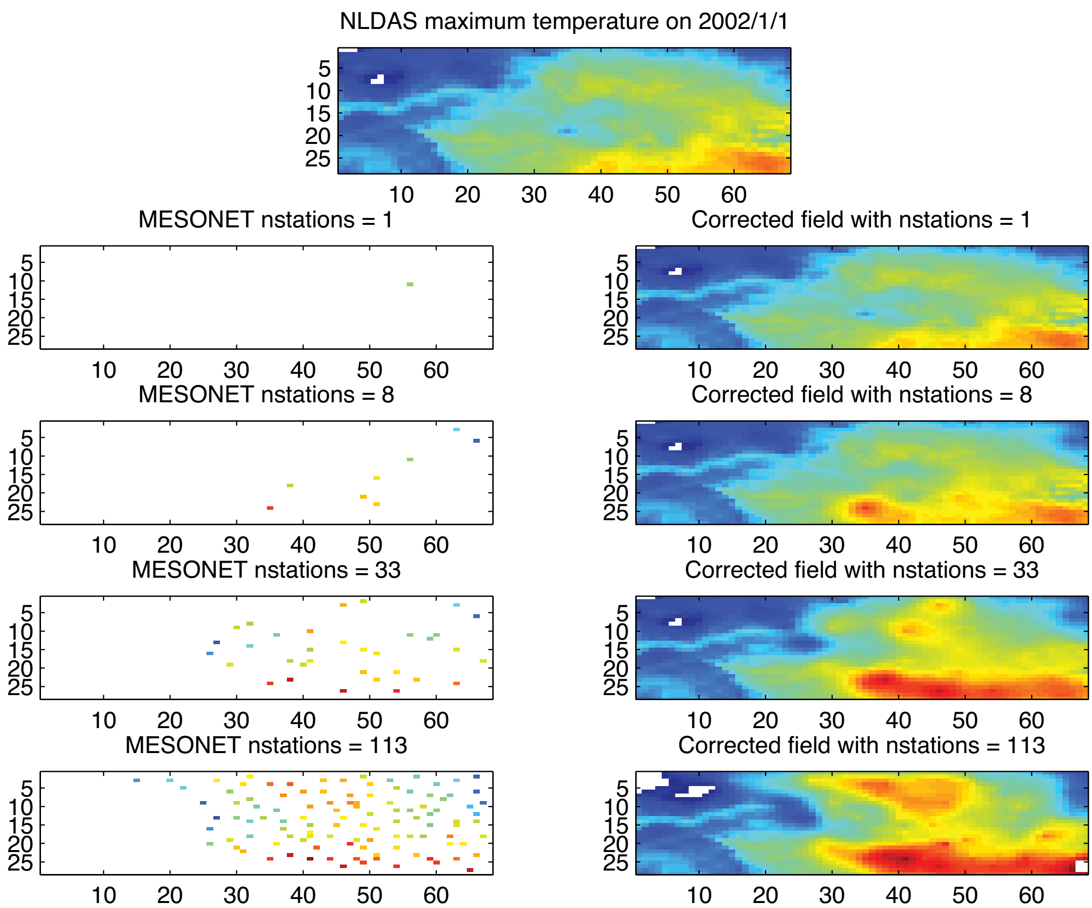


Figure 12 shows time series of the areal averages and the spatial variability over the OK Mesonet of maximum and minimum daily temperature and wind speed. The original NLDAS-2 data (blue) were corrected using data from all OK Mesonet stations. The areal mean values for temperature are not very different between the original and corrected datasets; this is expected because the NLDAS-2 dataset are generally reasonable for temperature over large scales. The spatial variability, however, increases in the corrected dataset because higher or lower values at individual stations are not replicated in the original dataset. The differences are much larger for wind speed, especially for the spatial variability. This is expected because the NLDAS-2 wind speed is based on the North American Regional Reanalysis dataset, which fails to represent the high spatial variability in wind speed.

Figure 13 shows the error in the merged dataset (averaged over all stations) when data from an increasing number of randomly selected stations are assimilated. Clearly with one station the correction is minimal. However, as the number of stations increases, the corrections are more spatially extensive and the errors are reduced. Increasing the number of stations from 1 to 20 has a large impact on accuracy, after which additional stations have little impact. This suggests that our method will have a positive effect on the increase in accuracy, even in regions with sparse networks such as in Africa.

2.5.2 Demonstration for Africa

This assimilation methodology was applied initially to the data set over the entire domain and one year (2000). Figure 14 shows an example of combining the gridded data set with the observations of daily maximum temperature. In regions where the original gridded data set is smooth, the spatial correlation is very high and the influence of the stations on these grid cells is large. In other regions, such as the East African Rift, the spatial correlation is low and the measurements influence grid cells only in their immediate vicinity.

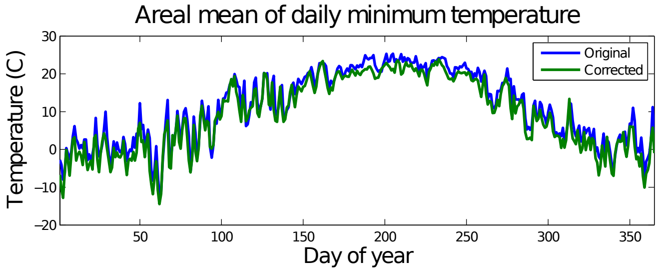
Figure 15 shows the effect of the assimilation procedure for 1 February 2000 for a region in Africa (Republic of the Congo) where the correlation between the observations and the original gridded data set, as shown in figure 6a, is low. The assimilation incorporates the observed lower maximum daily temperatures along the border between Congo and Gabon directly into the gridded data set and uses the existing spatial structure to influence the surrounding area.

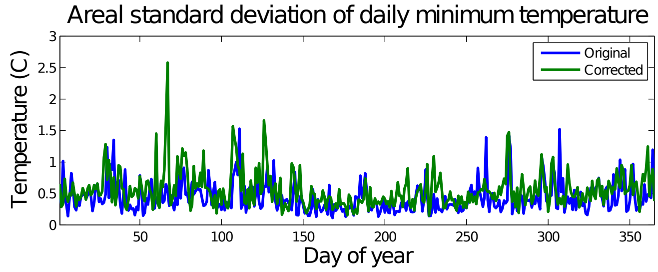
Figure 16 illustrates the impact of the assimilation on the mean and spatial variability of the Congo–Gabon region during 2000. The seasonality in both the mean and the spatial variability is maintained. The largest changes are an overall increase in spatial variability and a correction of the areal mean daily fluctuations to more closely match those of the observations.

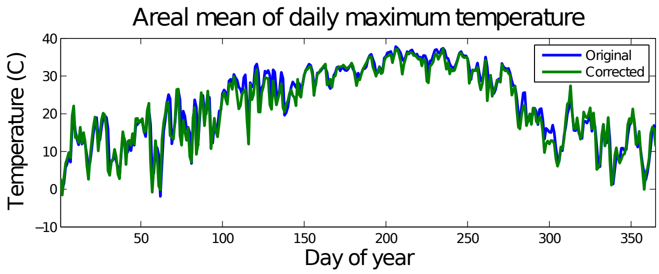
2.6 Implement existing methods to develop a 30-year meteorological dataset

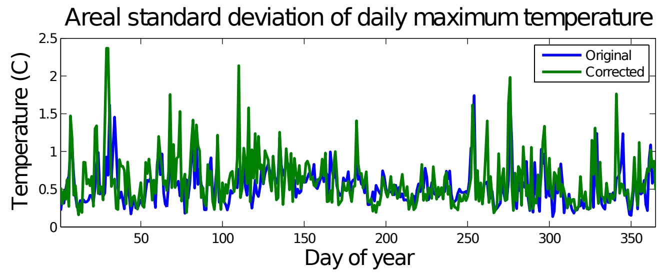
Building on Task 3, the methods of Sheffield and others (2006) were used to extend the initial 1990–2005 demonstration dataset to the full version for 1979–2008 at daily and 10 km resolution, covering 20°W to 60°E and 5°S to 25°N.

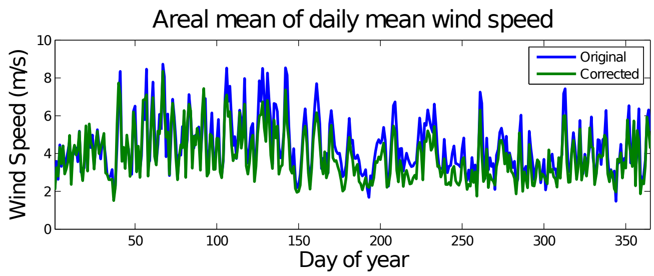
Figure 12. Comparison of the areal means and areal standard deviations over Oklahoma before and after the correction using the entire MESONET network with data during 2002











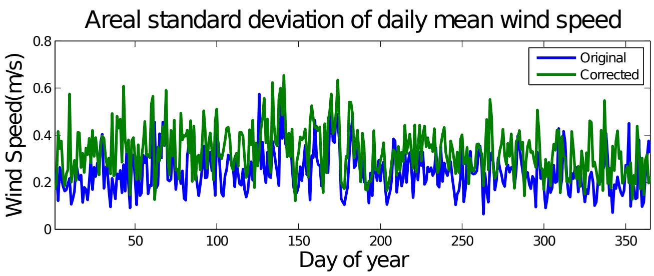


Figure 13. Absolute error in the merged daily maximum temperature dataset when in situ measurements from an increasing number of randomly selected stations from MESONET are assimilated into the background NLDAS fields over Oklahoma on 1 February 2002. The baseline for comparison is the merged field using all available MESONET stations.

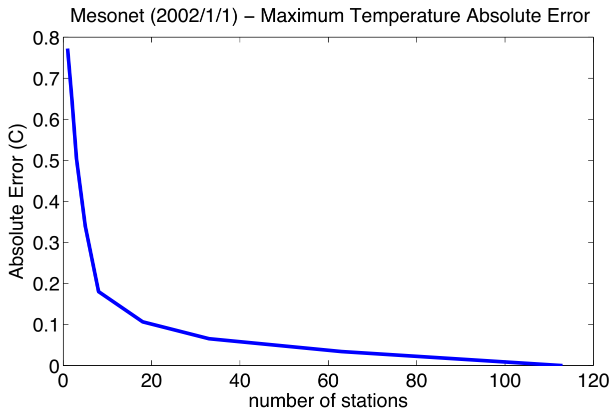


Figure 14. Merging the original gridded data (background field) with the observations (GSOD) to arrive at a corrected field of daily maximum temperature (K) for 1 February 2000

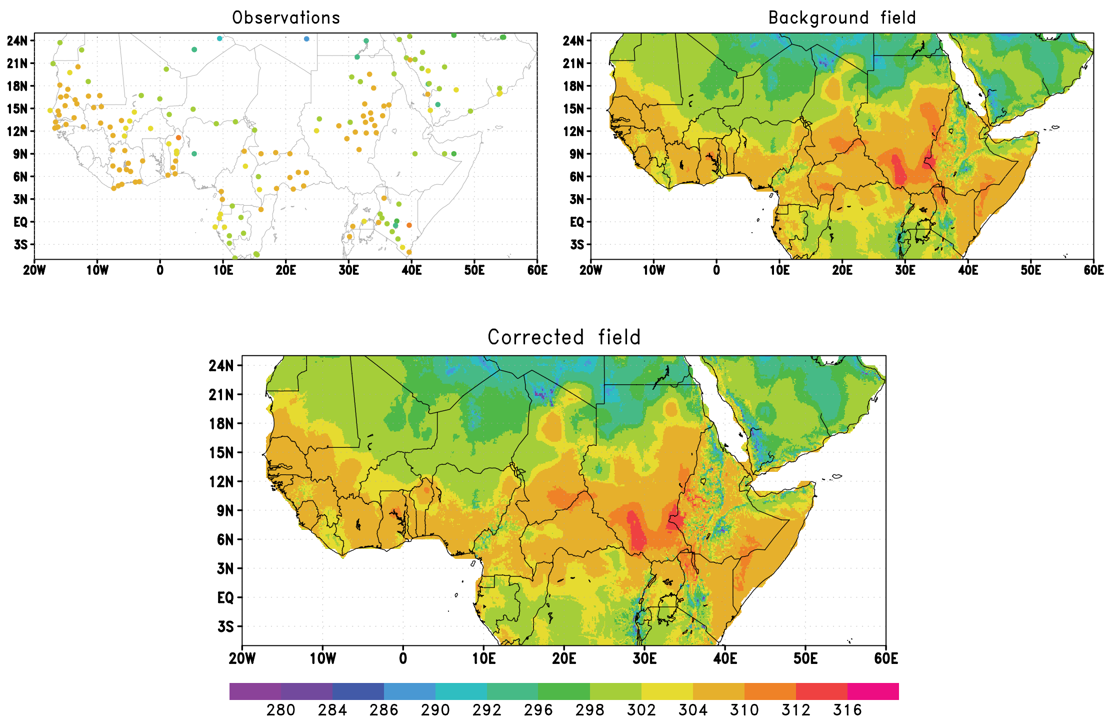


Figure 15. Original gridded data (background field) merged with the observations (GSOD) to arrive at a corrected field of daily maximum temperature (K) over the Congo/Gabon region for 1 February 2000

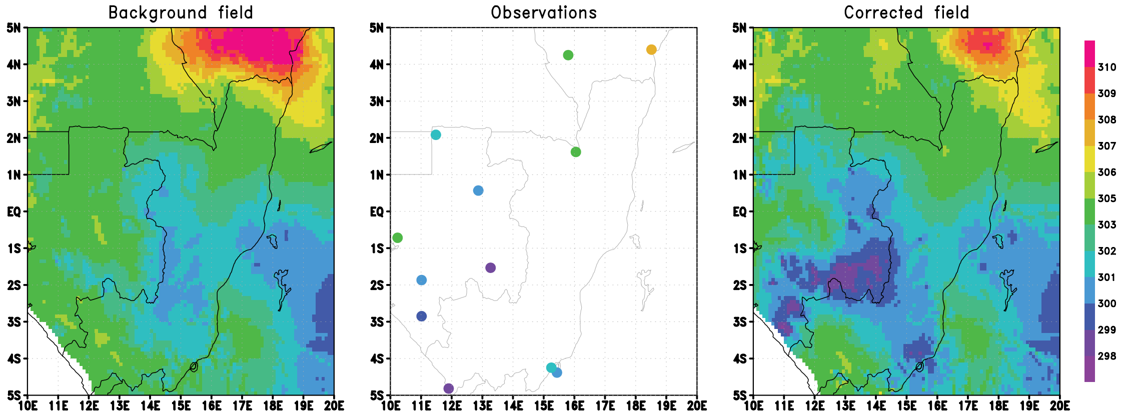
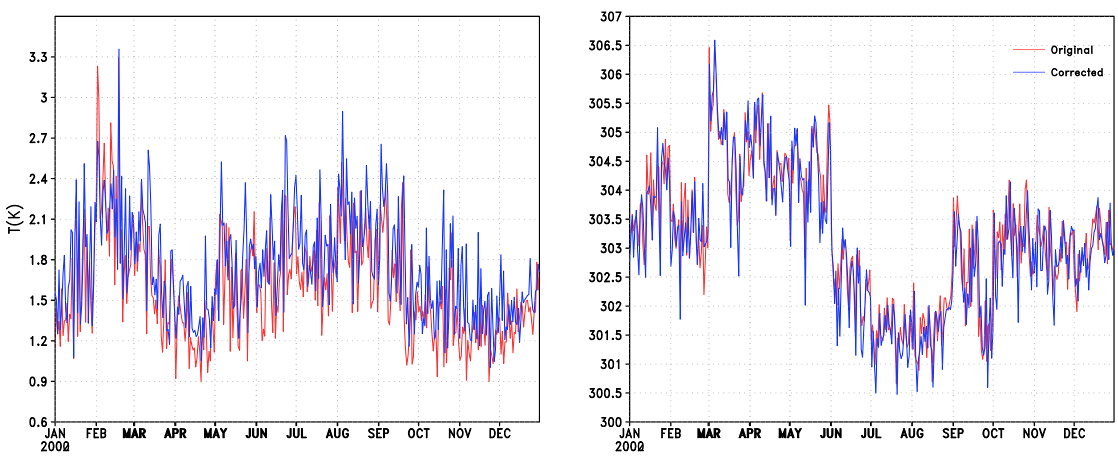


Figure 16. Comparison of the original and corrected gridded data set in the Congo/Gabon region for areal mean (right) and areal standard deviation (left) of daily maximum temperature during 2000



2.6.1 Infilling of missing station data to ensure temporal consistency

The assimilation method was applied to merge the GSOD station observations with the gridded dataset for the full period 1979–2008. This presented a number of challenges that become more apparent when merged over the full time period, especially for ensuring temporal consistency. Figure 17 shows that the vast majority of stations over the domain have data for less than 50% of days between 1979 and 2008. Excluding all stations that have less than 90% of temporal coverage results in a very small number. To ensure temporal consistency in the merged dataset, we required stations to have at least 20% of days with data, which precludes many and thus reduces the spatial coverage of stations. We therefore chose to implement an infilling procedure for missing GSOD station records that have more than 20% temporal coverage (figure 17) for the entire time period. This gave a better balance of temporal versus spatial coverage of the station data, assuming the errors in the infilled time series were reasonably low. The infilling procedure is as follows.

Assume the closest grid cell from a station acts as a proxy for the station itself. If a station value is missing, the value of the grid cell was bias-corrected using a relationship derived from the coinciding station and grid cell time-series when station data are available. This assumes that the daily variability in the gridded time series is reasonable. This results in a temporally consistent time series for 1979–2008 for each station. The bias correction is based on a quantile matching technique that translates the gridded value to an equivalent station value by searching for the equivalent quantile value in their respective empirical Cumulative Distribution Functions (CDF). The empirical CDF for the station and the grid cell were calculated from the daily data for all years using a 21 day window centred on the day of interest. This ensures that the corrected data are consistent with the distribution of the station data, thus correcting for the mean and variance.

Figures 18–20 illustrate the application of the method to three stations that have at least a 20% temporal coverage of daily data from 1979 to 2008. The available station data from the entire period is used to correct the bias in the grid-cell data for daily maximum temperature, minimum temperature, wind speed and dew point; this becomes a proxy time-series for the station data that infills the missing records. The technique effectively captures the seasonality and variance of the station data and was used to generate a station dataset of daily maximum temperature, minimum temperature, average wind speed and dew point for the full time period.

2.6.2 Assimilation of station data into the gridded dataset for 1979–2008

The infilled station data were then merged with the original gridded data to give a temporally consistent final dataset. An example of the influence of the station data on the gridded dataset is shown in figure 21 for 1 January 1979. A value of 1 equates to minimizing the mean squared error by relying exclusively on a linear combination of station bias-correction values while discarding the mean value. The closer the sum of weights is to zero, the more weight is given to the mean and the less to station values.

Where the network density is high in West Africa, as shown in figure 21, nearly all grid cells are adjusted with station information. In this case, the merging method is able to optimally interpolate the bias-corrected values. On the other hand, in countries such as Nigeria, where there are no stations with sufficient records, no merging is done.

Figure 17. Stations from the GSOD network that fulfill the requirements to be assimilated into the gridded data set are shown in green (> = 20% of days with data) while those that do not are in red (< 20%).

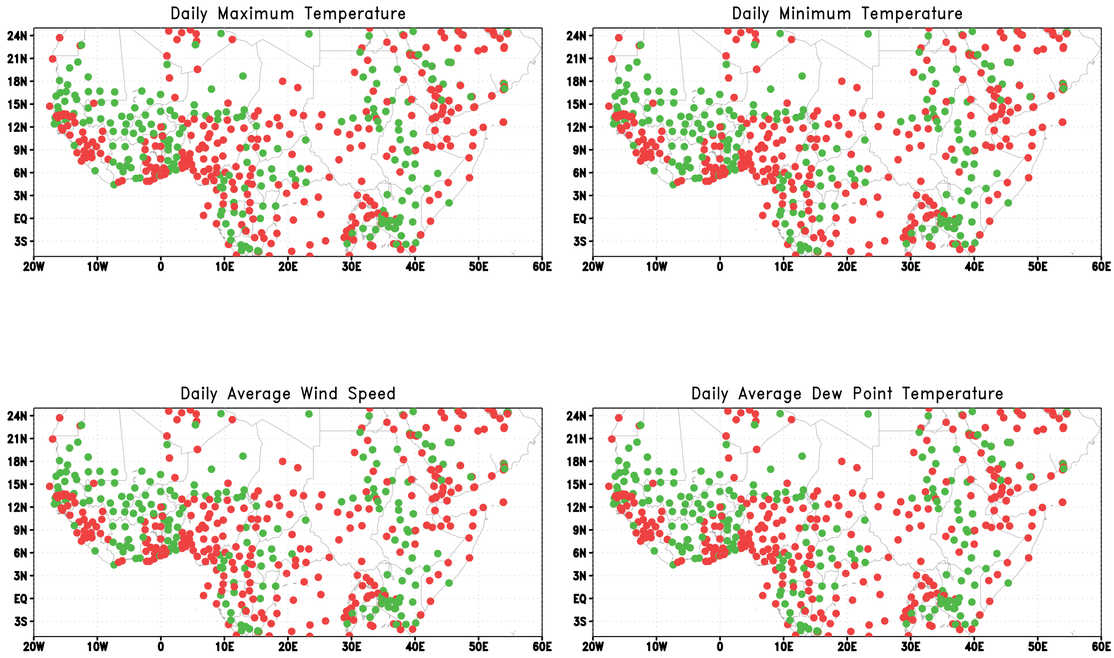


Figure 18. Example of infilling missing station records for station 624140. Data from the closest grid cell (blue) is used as a proxy to simulate the station data (red) that is missing while setting the original grid cell data to be equal to the station value. The grid cell values are bias corrected (green) against the existing station data using a quantile matching technique.

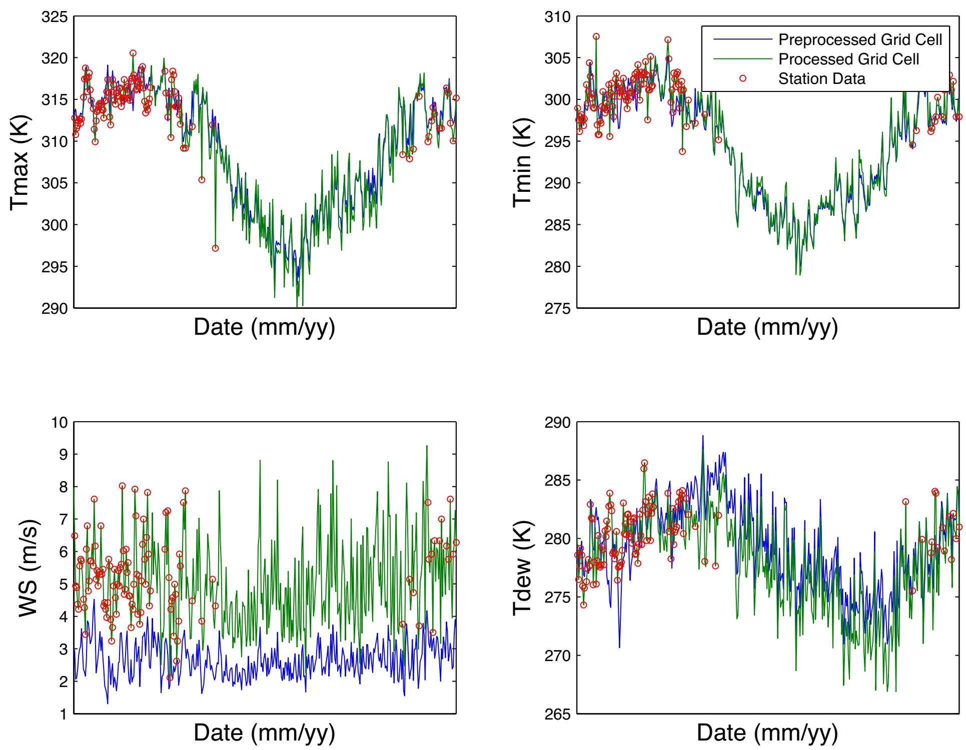


Figure 19. Same as figure 18, but for station 628050

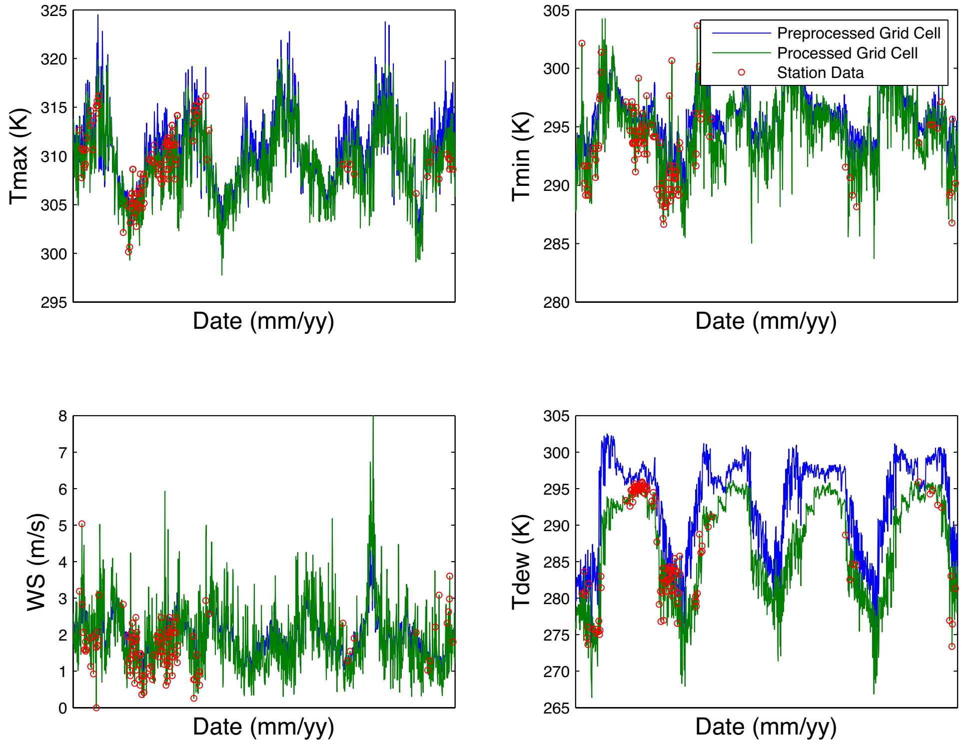


Figure 20. Same as figure 18, but for station 612850

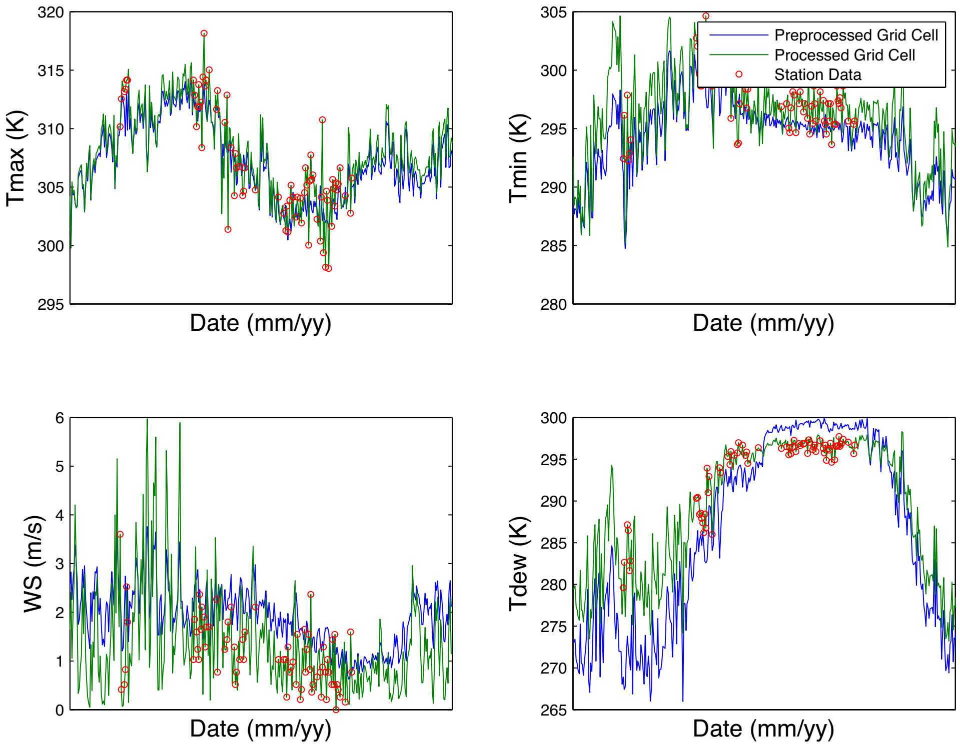
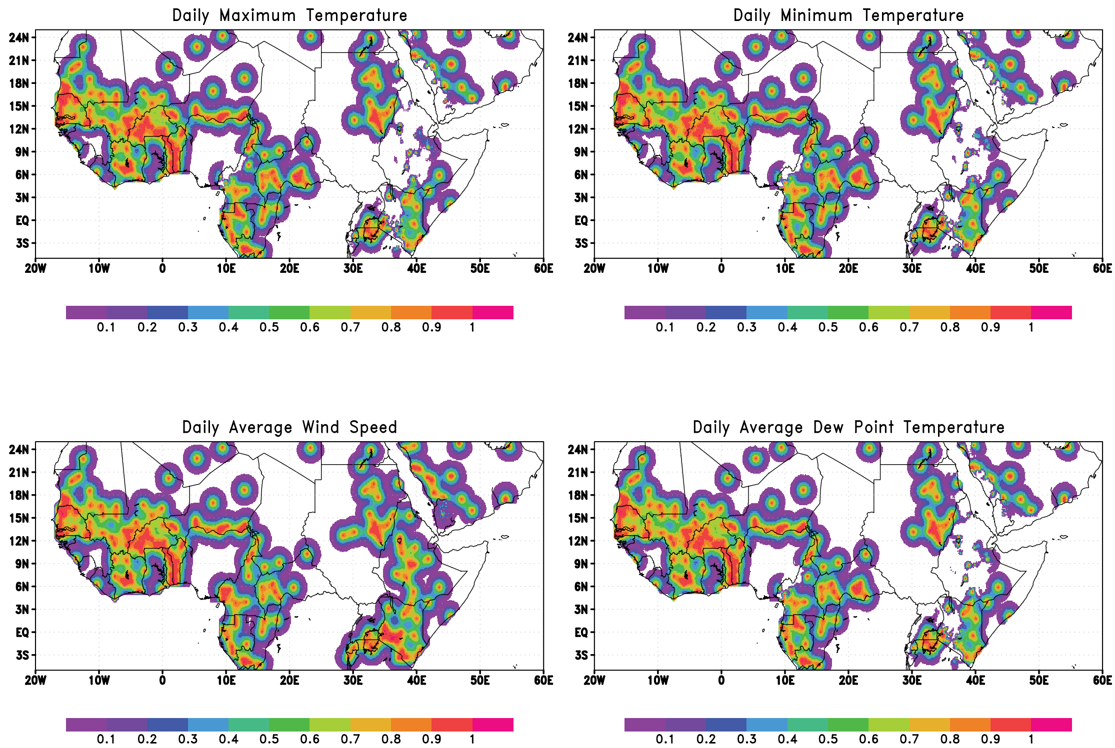


Figure 21. The sum of weights show the impact of stations on the interpolated bias correction values on 1 January 1979. Note that for computational reasons, a station is never allowed to influence more than 1 degree away, regardless of the spatial correlation.



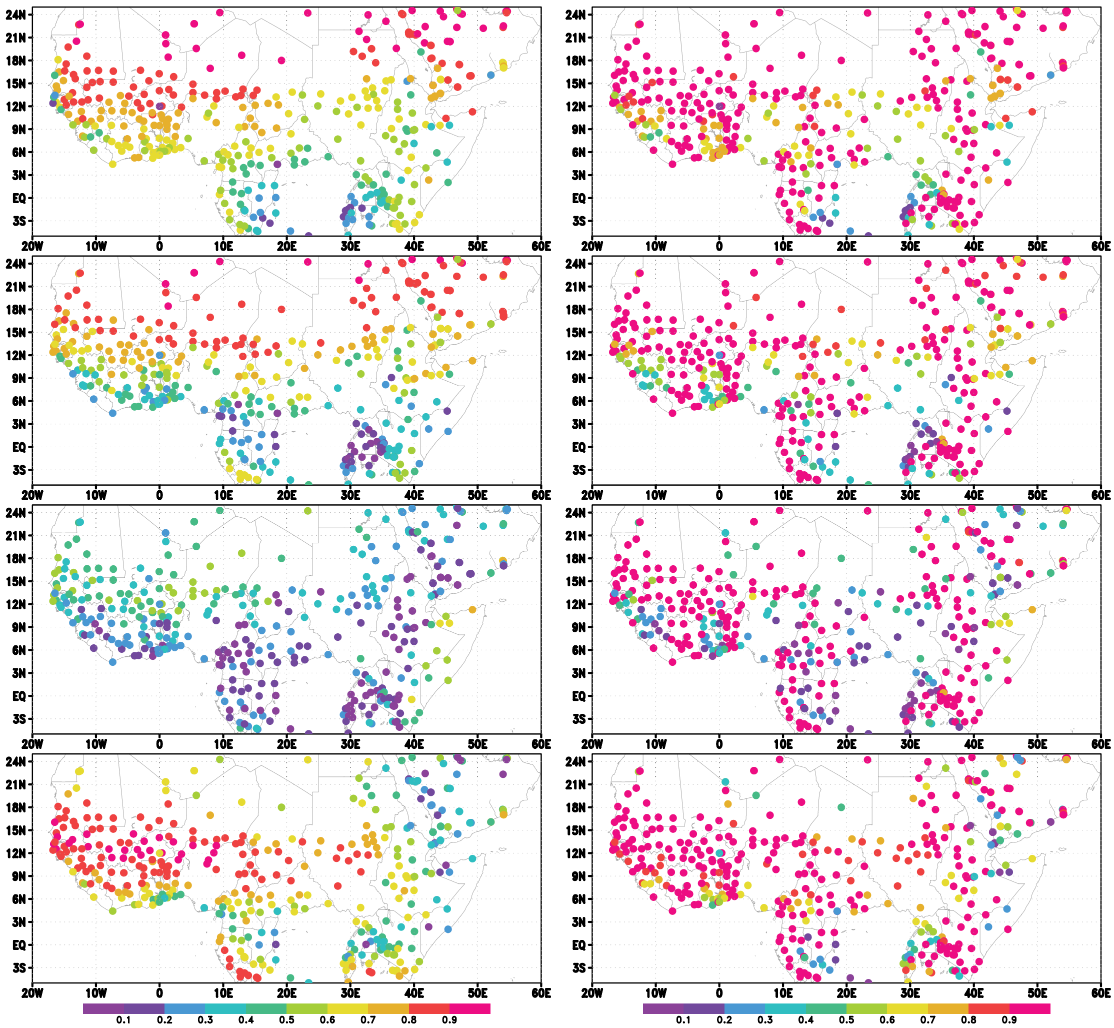
Due to the low density of the station networks, the spatial correlation of the gridded data set is used as a proxy for the spatial correlation of the biases. This results in low spatial correlations in the East African Highlands and therefore low impact on the gridded variables that are affected by topography. Because the original gridded dataset for wind speed was not scaled down to account for topography (or other factors), spatial correlation is still high. Future work should determine how to infer the spatial correlations of the biases from low-density networks.

The spatial-correlation functions were found at each time-step for each grid cell. This results in slightly different maps of the sum of weights of the bias-correction values. However, because all stations reported during the entire record (due to infilling), the variance and mean of the sum of weights were conserved. This ensures a temporally consistent final gridded product.

2.7 Evaluate the full dataset against station observations

Figure 22 shows how the Pearson correlation between the GSOD stations and their closest grid cells change after the assimilation. As expected, the grid cells that are closest to the stations used for the assimilation now show very high correlation values. The figure also shows the new correlations with the stations that were not used for the assimilation. In general, when the grid cell that corresponds to the station that was not used is close to another station that was used, there is an overall increase in correlation. When a station is

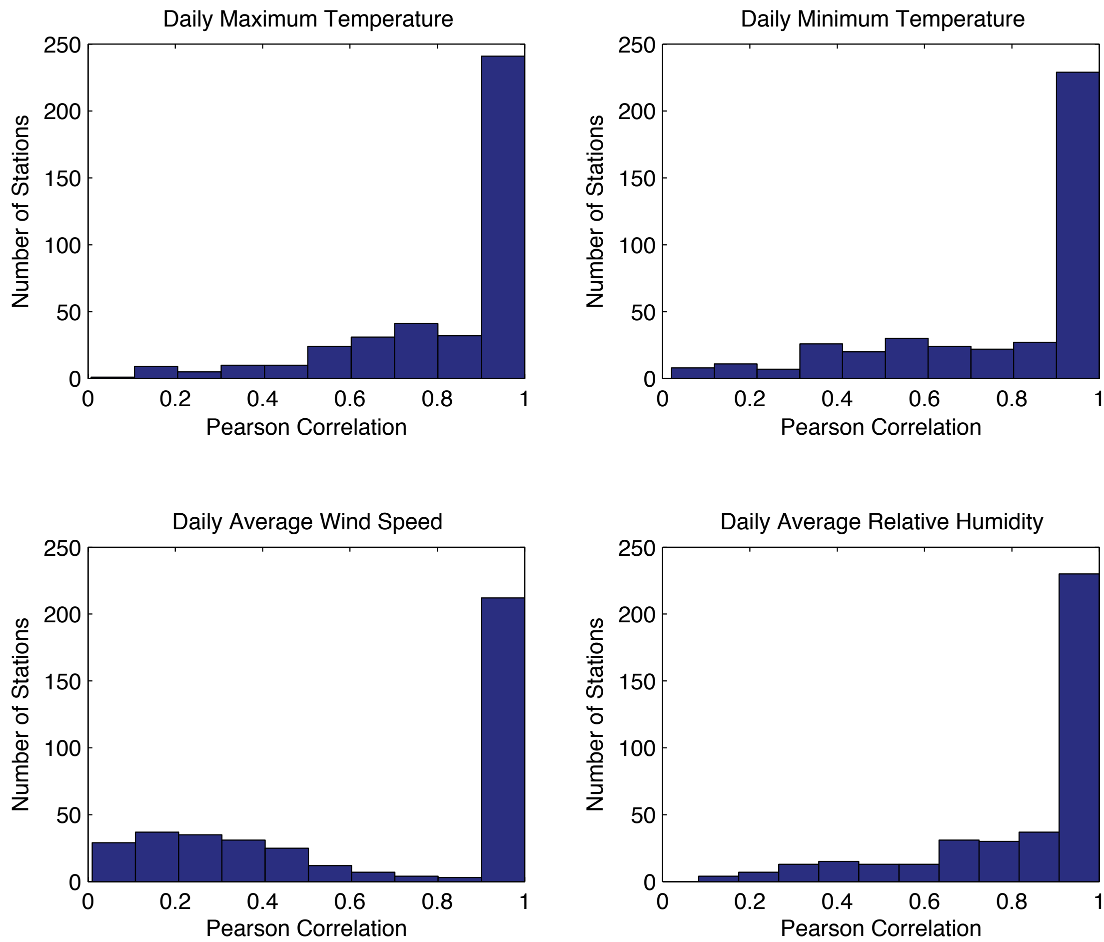
Figure 22. Comparison of the Pearson correlation of the closest grid cells to the GSOD stations before (left) and after (right) the assimilation procedure. The variables that are assimilated are daily maximum temperature (top), daily minimum temperature (top centre), daily average wind speed (bottom centre) and daily mean dewpoint (bottom). Note: If a station variable has less than 100 values, it is not used for this comparison.



almost entirely out of the influence of others that were assimilated, the increase in correlation is negligible. Figure 23 shows a histogram of the values of the Pearson correlations from figure 22. Except for average wind speed, all the variables now have an average Pearson correlation above 0.8.

In figure 24, three stations and their closest grid cells are chosen to illustrate the effects of the assimilation method at the level of grid cells. When the station is located at the centre of the grid cell, the corrected grid cell matches the station data perfectly (the uppermost graph in figure 24). When the station is slightly further away from the grid cell, the correlation is high but less than 1 because the station influence on bias correction decreases as a function of distance. The remaining two graphs in figure 24 are cases in which the stations do not have a

Figure 23. Histograms of Pearson correlation coefficients between the high-resolution gridded data and the GSOD station database for daily data



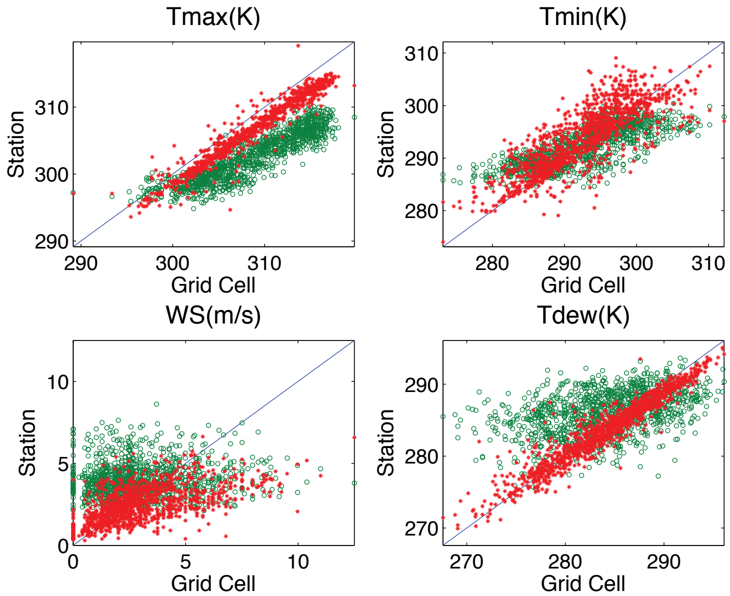
20% temporal coverage over the 1979–2008 period and thus were not used for the merging algorithm. These validate the algorithm's ability to correct grid cells when no station data is available.

When the influence of neighbouring stations is high, as defined by the spatial correlation, the station data correct the grid cell data to more closely match reality (middle graph in figure 24). The same is not true when the spatial correlation limits the station influence. In the limiting (or worst) case, where the influence of station data is 0, the grid-cell value does not change (figure 24, lower graph) and the correlations are very low, especially for wind speed.

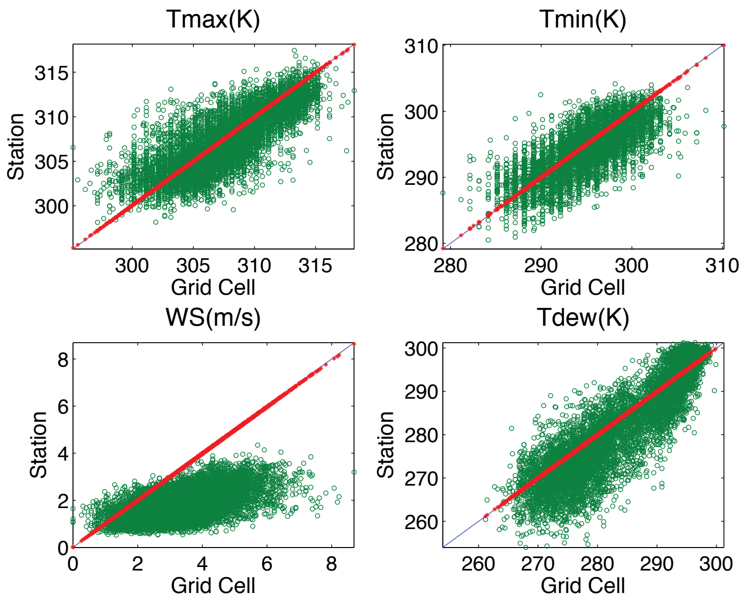
A comparison between spatial maps of the original and corrected gridded data sets is shown in figure 25. The annual mean of all the meteorological variables is computed for 1984. This year is relevant due to the severe drought in the Sahel. The corrected maximum temperature shows a lower daily maximum over the Sahel and a distinctively higher minimum temperature compared to the original gridded data. Also, higher dewpoints are found further north in the corrected dataset. The influence of isolated station data can be seen particularly in the northern Sahel for wind speed and these 'bulls-eyes' have been removed by increasing the spatial correlation length and rejecting isolated stations very different from the background field.

Figure 24. The original (green) and merged (red) grid cell data are compared to their closest station. These show the impact of having a station in the grid cell (top), close to the grid cell (middle) and far away from the grid cell (bottom).

Station number: 612910 - Latitude: 12.5330 Longitude: -7.9500

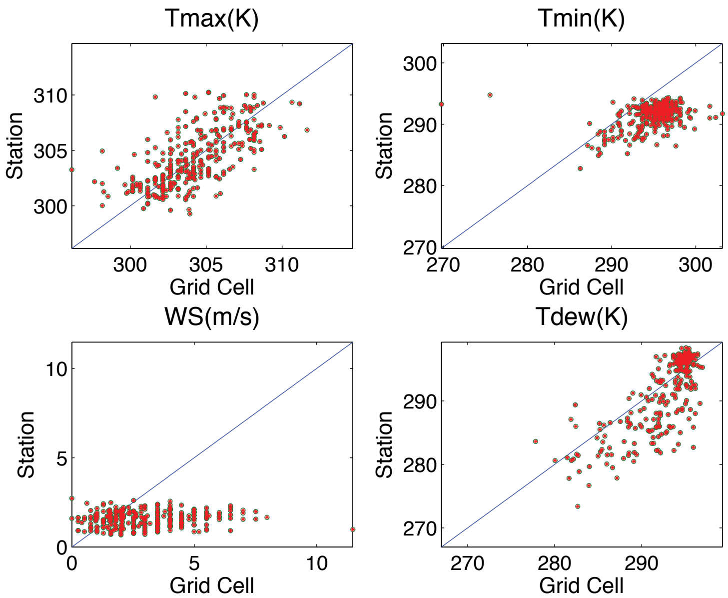


Station: 612910	Tmax	Tmin	WS	Tdew
Station vs original (Pearson ρ)	0.8002	0.7281	0.5897	0.9209
Station vs merged (Pearson ρ)	1.0	1.0	1.0	1.0



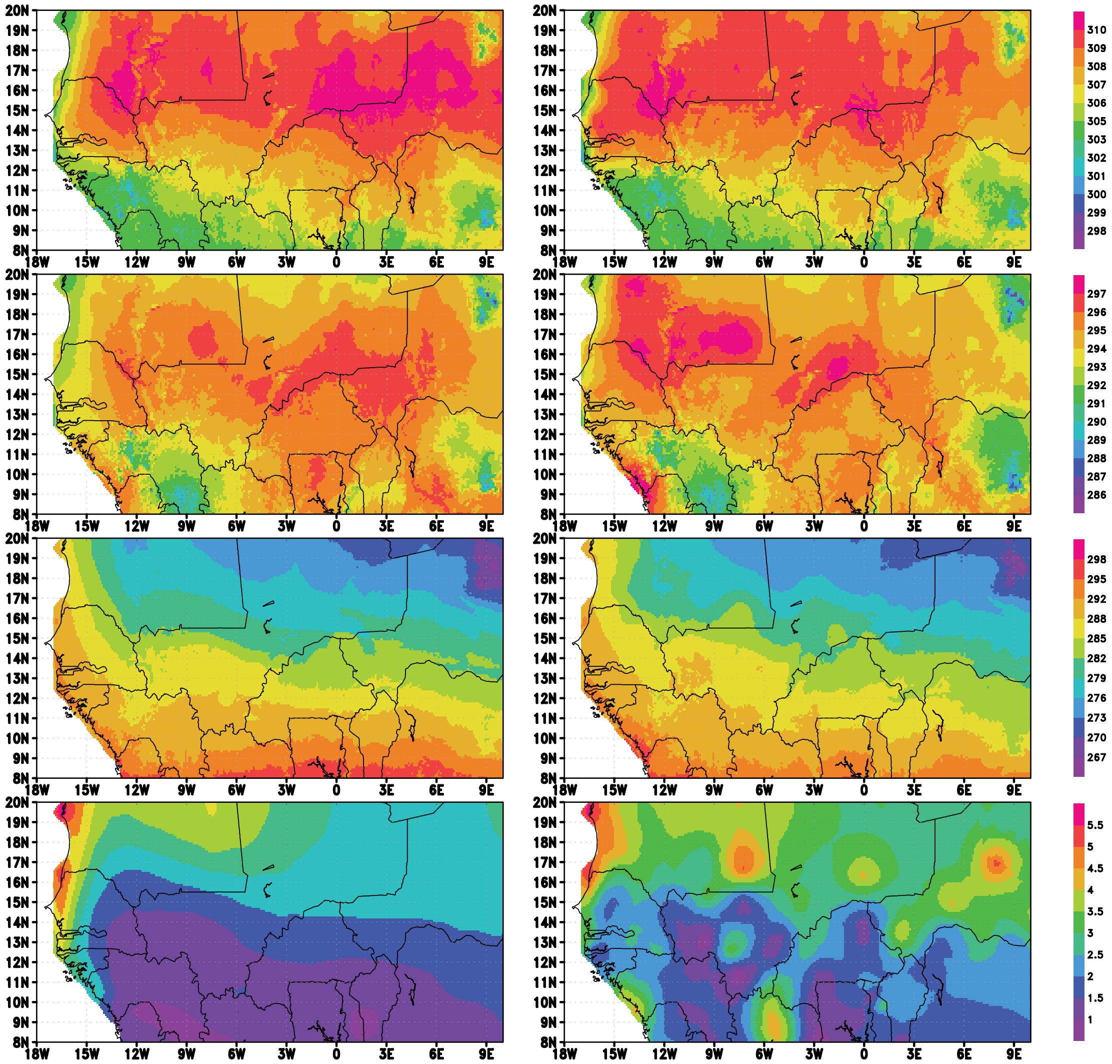
Station: 624140	Tmax	Tmin	WS	Tdew
Station vs original (Pearson ρ)	0.8961	0.7683	0.0651	0.4335
Station vs merged (Pearson ρ)	0.9730	0.8197	0.5815	0.9564

Figure 24. The original (green) and merged (red) grid cell data are compared to their closest station. These show the impact of having a station in the grid cell (top), close to the grid cell (middle) and far away from the grid cell (bottom) *(continued)*



Station: 618340	Tmax	Tmin	WS	Tdew
Station vs original (Pearson ρ)	0.6698	0.3662	0.0802	0.7732
Station vs merged (Pearson ρ)	0.6698	0.3662	0.0802	0.7732

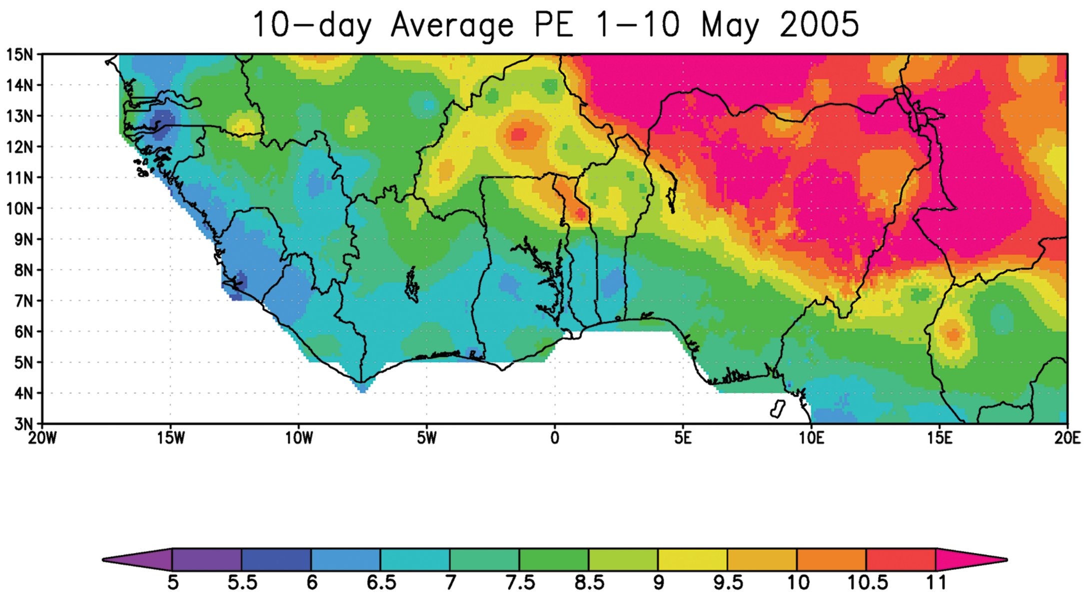
Figure 25. The original (left) and corrected (right) data sets are compared through their annual mean (1984) of the 4 daily meteorological variables: maximum temperature (top), minimum temperature (top centre), dewpoint temperature (bottom centre) and mean wind speed (bottom).



2.8 Development of value-added products

We have created a number of value added products including annual, monthly and daily statistics of the meteorological variables as well as characterization of extreme values. The latter includes a set of indices to characterize extreme daily values as defined by the Expert Team on Climate Change Detection and Indices (ETCCDI). These include the number of tropical nights (Tmin > 20°C) and duration of warm spells. We also generated a dataset of daily potential evaporation based on the Penman Monteith model forced by the merged dataset. This shows the utility of the dataset and is a precursor to its use in modeling crops. Figure 26 is an example.

Figure 26. Snapshot of 10-day average potential evaporation (mm/day) for West Africa calculated using the Penman-Monteith model forced by the meteorological dataset



2.9 Paper writing

A manuscript is in preparation that documents the methods and datasets produced under this project and evaluates the final merged dataset in terms of the climatology and trends in extreme values (Chaney et al. 2012).

3. Conclusion and recommendations

This project developed a 10 km daily dataset of meteorological variables for 1979–2008 for West and East Africa, which is suitable for forcing crop and other types of terrestrial models. Using existing methods for creating global products, available data from global gridded monthly observations were merged with high temporal resolution reanalysis data and scaled down in space to 10 km resolution. Corrections were made for temporal and inter-variable consistency. The dataset was further improved by assimilating daily station data where available. The final dataset was used to calculate indices of extreme daily values and potential evaporation, in order to demonstrate its potential for use in climate change studies and for forcing crop models.

Further improvements can be made by using additional sources of observational data. In particular, the methods developed provide a framework for merging new station data, such as from local stations of regional African partners that are generally not readily available, and from newly developed, high spatial resolution satellite-based surface radiation products.

4. References

Challinor AJ, Wheeler TR, Slingo JM, Craufurd PQ, Grimes DIF. 2005. Simulation of crop yields using ERA40: Limits to skill and non-stationarity in weather-yield relationships. *J. Applied Meteorology*, 44:516–531.

Chaney NJ, Sheffield J, Villarini G. 2012. High Resolution Gridded Daily Meteorological Data for Africa: Dataset Development and Analysis of Trends in Means and Extremes. To be submitted to *J. Climate*.

Chirlin GR, Wood EF. 1982. On the relationship between kriging and state estimation. *Water Resour. Res*., 18(2), 432–438, doi:10.1029/WR018i002p00432.

Ngo-Duc T, Polcher J, Laval K. 2005. A 53-year forcing data set for land surface models. *J. Geophysical Research*, 110:D06116, doi:10.1029/2004JD005434.

Robock A et al. 2003. Evaluation of the North American Land Data Assimilation System over the Southern Great Plains during the warm season. *J. Geophys. Res.*, 108 (D22), 8846, doi:10.1029/2002JD003245.

Rodionov SN. 2004. A sequential algorithm for testing climate regime shifts. *Geophys. Res. Lett.*, 31, L09204, doi:10.1029/2004GL019448.

Sheffield J, Ziegler AD, Wood EF, Chen Y. 2004. Correction of the high-latitude rain day anomaly in the NCEP/NCAR reanalysis for land surface hydrological modeling. *J. Climate*, 17(19), 3814–3828.

Sheffield J, Goteti G, Wood EF. 2006. Development of a 50-year high-resolution global data-set of meteorological forcings for land surface modeling. *J. Climate*, 19(13), 3088–3111.

Matsuura K, Willmott CJ. 2010: Terrestrial air temperature and precipitation: 1900–2008 gridded monthly and annual time series. Version 2.01, Center for Climatic Research, University of Delaware. [Available online at http://climate.geog. udel.edu/;climate/html_ pages/Global2_Ts_2009/README. global_t_ts_2009.html.]

Xia Y, Mitchell K, Ek M, Sheffield J, Cosgrove B, Wood E, Luo L et al. 2011. Continental-Scale Water and Energy Flux Analysis and Validation for the North-American Land Data Assimilation System Project Phase 2 (NLDAS-2), Part 1: Intercomparison and Application of Model Products. *J. Geophys. Res.*, accepted.

CGIAR

RESEARCH PROGRAM ON
Climate Change, Agriculture and Food Security

CCAFS

The CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS) is a strategic initiative of the Consultative Group on International Agricultural Research (CGIAR) and the Earth System Science Partnership (ESSP), led by the International Center for Tropical Agriculture (CIAT). CCAFS is the world's most comprehensive global research program to examine and address the critical interactions between climate change, agriculture and food security.

For more information, visit www.ccafs.cgiar.org

Titles in this Working Paper series aim to disseminate interim climate change, agriculture and food security research and practices and stimulate feedback from the scientific community.

CIAT

Earth System Science Partnership

PRINCETON UNIVERSITY