2025

REPORT ON THE MERGING OF STATION OBSEVATION WITH REANALYSIS(ERA5) DATA

ISMAEL LUTTA

Table of Content

1.0 Introduction

The Greater Horn of Africa (GHA) region experiences varied vagaries of weather ranging from extreme temperatures, frequent floods and droughts that affect the lives and livelihoods of many vulnerable communities. In the past, National Meteorological and Hydrological Services (NMHSs) in the region have placed higher priority on monitoring rainfall variability, compared to temperature variability. However, temperature monitoring, including events such as heatwaves, is becoming increasingly important. Thus, accurate and timely 2m temperature information is urgently required for monitoring the extreme climatic events being experienced in the GHA region.

The GHA region temperature observations are either scarce or not exchanged. Consequently, the region lacks a validated gridded temperature dataset (temperature station observations merged with reanalysis data) with good observational coverage that is updated in near-real time which can be used for monitoring. The aim of this work was to investigate the gaps that exist in data and utilise proxy datasets to develop gridded regional temperature datasets to be utilized for climate monitoring.

This need arose from the recommendation actions on data services (DS2) and Monitoring (CM1) from the ClimSA RCC-IGAD evaluation report (Graham, 2023). The temperature station observations are sparse over many parts countries in the GHA region. Many climate studies are in agreement that temperature datasets form, a vital component of climate analysis and products generation for critical climate-weather related decision making for protection of life and property from frequent climate induced disasters. Due to the sparse network of observation in the region, the use of reanalysis data as proxy data can adequately address this problem. However, a detailed investigations of these reanalysis datasets capabilities in reproducing atmospheric processes and patterns of near-surface temperature at spatial, seasonal and sub-seasonal timescales is required before the data can be used for merging with station observation to develop a regional or national database.

Several studies have been done to evaluate global reanalysis datasets in reproducing air temperature fields (Herbach et al., 2020). Zhu et al. (2021), Yilmaz (2023), and Rakhmatova et al. (2021) compared ERA5 reanalysis with observation stations in Antarctic, Turkey, and Uzbekistan, respectively. Tarek et al. (2020) evaluated ERA5 against observations temperature performance in North America. Gleixner et al. (2020) evaluated the performance of ERA5 temperature and precipitation against observations over East Africa. The performance of ERA5 reanalysis data in reproducing the daily variability of 2-meter temperature and main trends has been found promising (Cavaleri et al., 2024).

2.0 Merging station observation with reanalysis

The reanalysis raw datasets, were downloaded covering the Great Horn of Africa region of the bounding box as shown below;

Table 1*: Bounding box used to download reanalysis datasets*

Minimum Longitude	21.5 degrees East
Maximum Longitude	52 degrees East
Minimum Latitude	-12 degrees South
Maximum Latitude	23.5 degrees North

The reanalysis of 2-meter temperature (ERA5, JRA3Q, MERRA2 and NOAA), extracted at the Meteorological and Hydrological Services (NMHSs) observation station network was evaluated based on a comparison with the long time series (1981-2020) stations observation. The results of the statistic metrics obtained revealed that, ERA5 is the best-performing reanalysis dataset in most parts of the GHA region, with the highest correlation, lowest errors, and smallest bias.

It is well known that, the measurement of a meteorological variable taken at meteorological observation stations represents the true value of a measured meteorological variable. However, due to the sparse station network in many parts of the GHA region, there is need for the use of additional proxy data such as reanalysis to provide measurements of meteorological variables, especially temperature in areas where there are no stations. The limitation of the station observation is worse in regions with complex varying topography, where stations' measurements are sparse or no-existent. Reanalysis datasets can provide an alternative climate data in regions where ground observations measurements are limited or unavailable. However, without direct reference to ground measurements, reanalysis datasets are subject to systematic biases. Using Climate Data Tools (CDT), the main processing steps in producing gridded(merged) datasets, include the following steps;

i. Computing downscaling coefficients,
ii. Downscaling data,
iii. Computing bias coefficients,
iv. Applying bias correction and finally
v. Merging adjusted reanalysis temperature with station observations.

Downscaling is the collective term for the methods used to regionalize information from Global Climate Model (GCM) at a coarser spatial resolution and create a higher spatial resolution data or a fine spatial scale data (ground station). Its purpose is to bring the GCM model data in closer agreement with the station level data. By the use of CDT, computing downscaling coefficients is done on the assumption that, as altitude changes, temperature also changes with height. Hence a constant lapse rate (vertical temperature gradient) is computed for each month using station temperature data and elevation data from a digital elevation model. These coefficients are applied to the reanalysis data using a linear model to produce downscaled data. The reanalysis data, at 25 by 25 km resolution was corrected using the Elevation data (Digital Elevation Model, at 5 by 5 km resolution) to give the Downscaled Era5 Reanalysis at fine resolution of 5 by 5 km as shown in the Figure 2 below;

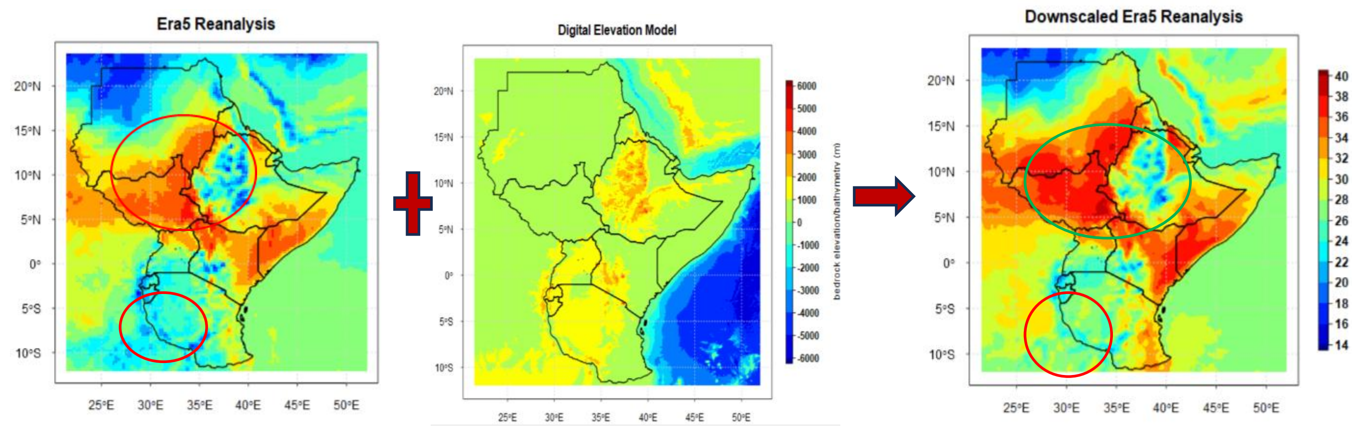

Figure 2: *Illustration of the spatial downscaling method for maximum temperature data.*

Once the downscaling was completed, the step that follows was to correct the bias of the gridded data, namely the reanalysis downscaled temperature data. There are many ways to correct specific errors and artifacts in the gridded data. It can be corrected regardless of the source of the error using ground truth measurements such as station data.

A multiplicative model was used to correct the bias in the downscaled ERA5 maximum and minimum temperature data. Climatological mean bias factor was computed for each dekad per year at the dekadal time step of the ERA5 reanalysis downscaled data. This was done to quantify the bias of the gridded data at the station locations, with assumption that the station data represent the true measurements of the maximum and minimum temperature values recorded at the stations.

First, the gridded data is extracted at the station locations, then the series pairs (station and extracted data) are used to compute the bias factors. Then, the bias factors are interpolated on downscaled temperature data using inverse distance weighted interpolation. A global interpolation is used. Table 2 present the parameters used to compute the bias factors for ERA5 temperature data. For dekadal data 36 bias factors are computed. To correct the bias, the ERA5 downscaled maximum and minimum temperature data are multiplied by the interpolated bias factors coefficients to adjust the downscaled reanalysis datasets.

Table 2: *Parameters used to compute the mean bias factors for JRA-55 downscaled maximum and minimum temperature data*

Bias computation method	Multiplicative Bias for Each Month
Base period	1981-2020
Minimum number of years of available data	15
Interpolation method	Inverse Distance Weighted (IDW)
Minimum number of stations to interpolate a grid	3
Maximum number of stations to interpolate a grid	9
Maximum radius of influence	4.0 degree

Finally, CDT has four methods of merging (Simple Bias Adjustment (SBA), Cressman Scheme (CSc), Barnes Scheme (BSc) and Regression Kriging (RK). See CDT [Merging Data](). Simple Bias Adjustment was used to merge station data with reanalysis corrected datasets. The selection of the merging and interpolation methods used to generate regional 2-meter temperature datasets, were informed by the previous work done by Faniriantsoa, R. (2023), Updated ENACTS Data for Kenya Meteorological Department.

CDT allows one to perform a multiple merging in a nested way, the outputs of the previous merging are used as input for the current merging process. Multiple passes are made through the grid at consecutively smaller radius of influence and number of neighbor stations used to increase precision. At each pass, the radius of influence, minimum and maximum number of neighboring stations used to interpolate on grid node are decreased in a given ratio. The table 3 below summarizing the process of merging with 4 nested runs.

Table 3*: Merging parameters used*

Run number	Pass ratio	maxdist	nmin	nmax	Input data	Output data
1	1	4	8	16	Initial input (E.g.: Bias adjusted temp gridded data)	Output 1
2	0.75	4 * 0.75 = 3	8 * 0.75 = 6	16 * 0.75 = 12	Output 1	Output 2
3	0.5	4 * 0.5 = 2	8 * 0.5 = 4	16 * 0.5 = 8	Output 2	Output 3
4	0.25	4 * 0.25 = 1	8 * 0.25 = 2	16 * 0.25 = 4	Output 3	Output 4

Where, nmin and nmax are the minimum and maximum number of neighboring stations used to interpolate on grid node respectively. Maxdist is the maximum distance in degrees. See CDT https://iri.columbia.edu/~rijaf/CDTUserGuide/html/merging_methods.html

3.0 Validations of merged datasets

Merged maximum temperature datasets were validated against the station observation from the period (1991-2020). This period was chosen as because it is the current baseline climatological period. CDT was used to validate multiple datasets, stations observations and extracted merged datasets at station location.

Figure 2a shows the scatter plots of dekadal maximum temperature from stations against the different merged data extracted at the station locations and the statistical indicators are shown in Figure 2b. The Simple Bias Adjustment merging method was used.

4.0 Results and Discussion

Figure 2a, the scatter plot shows that most data points are tightly clustered around the 1:1 line, showing a strong correlation and agreement with few outliers visible, indicating that the merged dataset performs well across a wide range of temperatures. Figure 2b, the Cumulative density Function (CDF) curves for both observed and merged values are nearly overlapping, indicating that the statistical distribution of the merged data closely matches the observed data.

Table 3, shows the statistical indicators for merged maximum temperature against the observations. The correlation metrics (CORR = 0.996) shows a strong positive linear relationship between observed and the merged data values. The Coefficient of determination (R2) multiplied by the regression slope (BR2 = 0.987) indicates both strong fit and close agreement. The bias (BIAS = 0.997) near 1, suggesting very little bias. The percent bias (PBIAS = -0.326) indicating a slight overestimation bias in the merged data. The Mean Error (ME= -0.11) and Mean Absolute Error (MAE = 0.354), both show small errors and average magnitude of errors in the merged data respectively. The Nash-Sutcliffe Efficiency (NSE = 0.991) shows how the merged datasets align with observations.

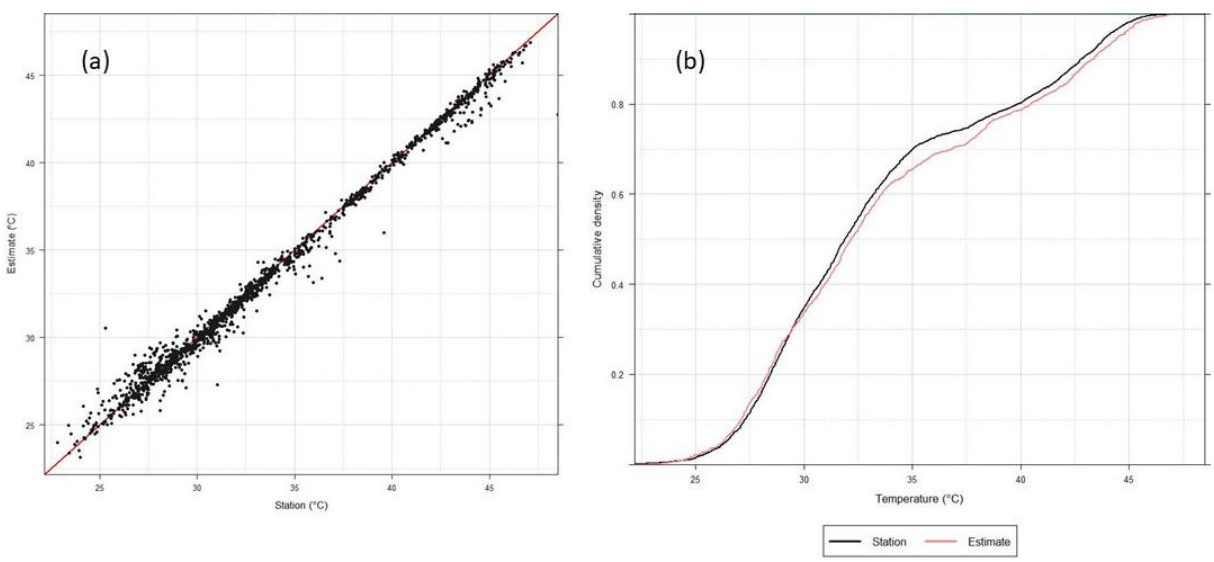

Figure 2 a, b*: Scatter plot and cumulative density Function plots for maximum temperature (1991-2020), stations against merged*

Table 4*: statistic indicators for the merged maximum temperature*

Name	MERGED TMX	Description
CORR	0.996	Correlation
BR2	0.987	Coefficient of determination (R2) multiplied by the regression slope
BIAS	0.997	Bias
PBIAS	-0.326	Percent Bias
ME	-0.11	Mean Error
MAE	0.354	Mean Absolute Error
RMSE	0.581	Root Mean Square Error
NSE	0.991	Nash-Sutcliffe Efficiency

5.0 Conclusion

The statistics demonstrate that the merged data closely aligns well with station observations having a high correlation (CORR) and Nash-Sutcliffe Efficiency (NSE), low Mean Absolute Errors (MAE) and Root Mean Square Errors (RMSE), and minimal bias hence a close match to observation data.

Reference

Cavalleri, Francesco, Francesca Viterbo, Michele Brunetti, Riccardo Bonanno, Veronica Manara, Cristian Lussana, Matteo Lacavalla, and Maurizio Maugeri. "Inter-comparison and validation of high-resolution surface air temperature reanalysis fields over Italy." *International Journal of Climatology* 44, no. 8 (2024): 2681-2700.

Faniriantsoa, R. (2023). Updated ENACTS Data for Kenya Meteorological Department.

Gleixner, S., Demissie, T. & Diro, G.T. (2020) Did ERA5 improve temperature and precipitation reanalysis over East Africa? Atmosphere, 11(9), 996. Available from: https://doi.org/10. 3390/atmos11090996

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horanyi, A., Muñoz-Sabater, J. et al. (2020) The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 146(730), 1999–2049. Available from: https://doi.org/10.1002/qj.3803

Rakhmatova, N., Arushanov, M., Shardakova, L., Nishonov, B., Taryannikova, R., Rakhmatova, V. et al. (2021) Evaluation of the perspective of ERA-interim and ERA5 Reanalyses for calcu lation of drought indicators for Uzbekistan. Atmosphere, 12(5), 527. Available from: https://doi.org/10.3390/atmos12050527

Yilmaz, M. (2023) Accuracy assessment of temperature trends from ERA5 and ERA5-land. Science of the Total Environment, 856 (January), 159182. Available from: https://doi.org/10.1016/j. scitotenv.2022.159182

Zhu, J., Xie, A., Qin, X., Wang, Y., Bing, X. & Wang, Y. (2021) An assessment of ERA5 reanalysis for Antarctic near-surface air temperature. Atmosphere, 12(2), 217. Available from: https:// doi.org/10.3390/atmos12020217