Are raw satellite bands and machine learning all you need to retrieve actual evapotranspiration?

Chouaib EL HACHIMI[1], *Saïd* KHABBA[1,3], *Salwa* BELAQZIZ[1,2], *Bouchra* AIT HSSAINE[1], *Mohamed Hakim* KHARROU[4], *Abdelghani* CHEHBOUNI[1,4]

[1]Center for Remote Sensing Applications (CRSA), Mohammed VI Polytechnic University (UM6P), Benguerir, Morocco.
[2]LabSIV Laboratory, Department of Computer Science, Faculty of Science, UIZ University, Agadir, Morocco.
[3]LMFE, Department of Physics, Faculty of Sciences Semlalia, Cadi Ayyad University, Marrakesh, Morocco.
[4]International Water Research Institute (IWRI), Mohammed VI Polytechnic University (UM6P), Benguerir, Morocco.

Abstract. Accurately estimating latent heat flux (LE) is crucial for achieving efficiency in irrigation. It is a fundamental component in determining the actual evapotranspiration (ETa), which in turn, quantifies the amount of water lost that needs to be adequately compensated through irrigation. Empirical and physics-based models have extensive input data and site-specific limitations when estimating the LE. In contrast, the emergence of data-driven techniques combined with remote sensing has shown promising results for LE estimation with minimal and easy-to-obtain input data. This paper evaluates two machine learning-based approaches for estimating the LE. The first uses climate data, the Normalized Difference Vegetation Index (NDVI), and Land Surface Temperature (LST), while the second uses climate data combined with raw satellite bands. In-situ data were sourced from a flux station installed in our study area. The data include air temperatures (T_a), global solar radiation (R_g), and measured LE for the period 2015–2018. The study uses Landsat 8 as a remote sensing data source. At first, 12 raw available bands were downloaded. The LST is then derived from thermal bands using the Split Window algorithm (SW) and the NDVI from optical bands. During machine learning modeling, the CatBoost model is fed, trained, and evaluated using the two data combination approaches. Cross-validation of 3-folds gave an average RMSE of 27.54 $W.m^{-2}$ using the first approach and 27.05 $W.m^{-2}$ using the second approach. Results raise the question: Do we need additional computational layers when working with remote sensing products combined with machine learning? Future work is to generalize the approach and test it for other applications such as soil moisture retrieval, and yield prediction.

1 Introduction

Evapotranspiration (ET) plays a crucial role as the junction between the energy balance and water balance components of the Earth's system. It represents the combined processes of water loss through evaporation from the Earth's surface and transpiration from vegetation. In

agriculture, accurately estimating actual evapotranspiration (ETa) is a key to achieving efficiency in irrigation. By harnessing ETa data and integrating it into irrigation management systems, farmers can make informed decisions that benefit both their bottom line and the environment[1][2]. Several empirical and physics-based models are being used to estimate the ETa[3][4]. These models still face weaknesses, especially in their reliance on extensive input data, the simplification of assumptions regarding complex processes, and the need for parameterization and model calibration using ground-based measurements when changing the study site[5]. Not to mention that these measurements may be sparse and scarce in some regions. These limitations restrict the spatial coverage capabilities of the models. As a result, there has been growing interest in leveraging Remote Sensing (RS) technologies combined with data-driven methods for ETa estimation as reported in [4].

Remote Sensing has the potential to deliver Earth observations from space, determining the state of the Earth remotely at any given time thanks to various instruments and sensors installed on different types of satellites covering different wavelengths of the electromagnetic spectrum. While RS satellites cannot directly provide estimates of ETa, most studies use a combination of derived parameters such as vegetation indices (VIs)[6][7][8][9][10][11], Land Surface Temperature (LST)[12][13], meteorological data[14][15], Land Use Land Cover (LULC), and Digital Elevation model (DEM), as input features for their data-driven models. However, the preprocessing steps needed to calculate the indices or retrieve parameters require additional computational power. This latter is a real constraint for deploying generalizable operational solutions. In addition, machine learning models are capable of discerning and acquiring the hidden complex patterns and relationships from the data autonomously, eliminating the need for extensive feature engineering, especially in supervised learning using structured data[16]. The question remains: Are raw satellite bands sufficient features to provide accurate ETa estimates? This research paper evaluates this approach in a rainfed wheat field in Morocco by investigating whether it can yield ETa estimates that are comparable in accuracy to traditional methods that use derived parameters.

2 Study Area and Data

2.1 Study Area

Sidi Rahal, our study area, is situated in the east Tensift Basin, central Morocco (Fig. 1). It exhibits a semi-arid Mediterranean climate, with an annual average precipitation of approximately 250 mm and an atmospheric evaporative demand (ET0) of approximately 1600 $mm.yr^{-1}$ according to the FAO method[17]. The selected study area was cultivated with wheat during the 2014-2015, 2016-2017, and 2017–2018 agricultural seasons, while it was left uncultivated (remaining bare soil) during the 2015-2016 season due to prevailing climatic conditions[18].

2.2 In Situ Data

An eddy covariance (EC) system was installed in the field at a height of 2 meters. This EC tower is equipped with a CNR1 radiometer (Kipp and Zonen) to measure the four components of net radiation (R_n) and is complemented by several heat flux plates (HFT3-L, Campbell Scientific Ltd) responsible for assessing soil heat flux (G). The analysis of energy balance closure revealed that the available energy ($Rn-G$) generally exceeded the EC measurements. In order to maintain energy balance, adjustments were applied to the sensible and latent heat fluxes (H and LE) using the Bowen ratio method[19]. Adjacent to the EC tower, a

meteorological weather station has also been installed to measure air temperature, solar radiation, relative humidity, wind speed, and rainfall at a half-hourly scale.

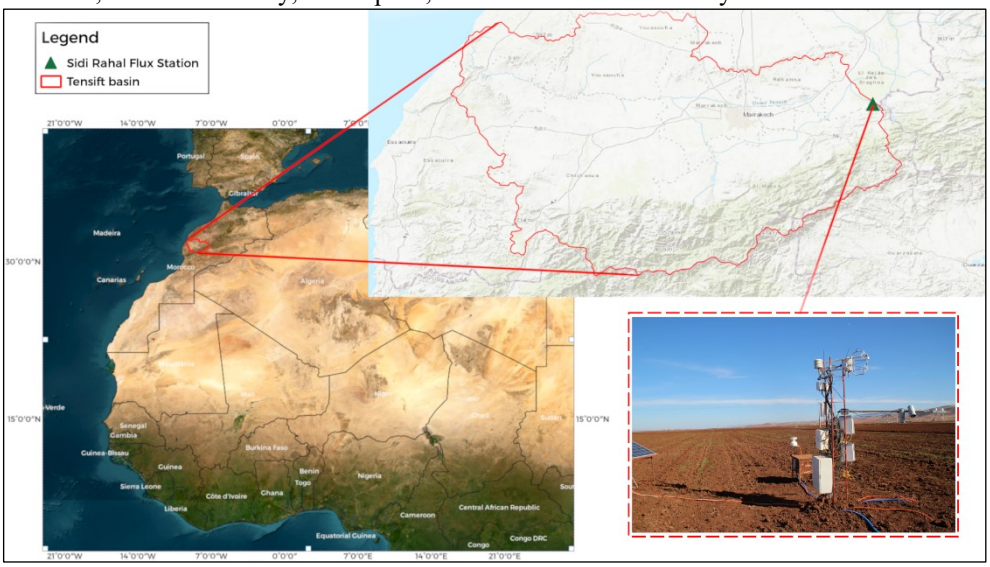

Fig. 1. Location of the Sidi Rahal rainfed wheat flux station in Tensift basin, Central Morocco.

2.3 Remote Sensing Data

The study relies on the Landsat, which is a joint program of the US Geological Survey (USGS) and NASA. It has been observing the Earth continuously since 1972 to present time. The satellites of this program cover the entire Earth's surface at a 30-meter resolution approximately every two weeks generating various types of data including multispectral and thermal data. The study uses Landsat 8 Collection 2 Tier 1 Digital Numbers (DN) values. These values represent a scaled, calibrated at-sensor radiance. Within this dataset, an array of 12 distinct bands of the electromagnetic spectrum is made available.

3 Methods

3.1 The CatBoost model

The high-performance gradient-boosting algorithm, known as CatBoost[20], combines multiple weaker models into a strong ensemble architecture that is efficient at handling different types of data and has internal mechanisms to manage several common problems in machine learning such as categorical features handling, and feature scaling. CatBoost outperformed the most popular open-source models such as XGBoost[21] and LightGBM[22] on several well-known machine learning tasks. It is designed to tackle the common issue of overfitting, a situation where a model becomes too tailored to the training data and loses its ability to generalize to new unseen data (testing or production data). The algorithm integrates mechanisms that automatically control and manage the complexity of the model, preventing overfitting and resulting in a more robust and reliable model performance. Furthermore, the algorithm has an integrated approach to address missing data, enabling it to work with datasets that might have incomplete information, without requiring extensive preprocessing to impute missing values.

3.2 Baseline approach

The study followed the steps shown in the flowchart of Figure 2. It consists of evaluating two input data combinations for LE estimation:

a. Using LST, Normalized Difference Vegetation Index (NDVI), and climate data

In this approach the features set contains LST, NDVI, T_a and R_g. First, the NDVI is calculated from the visible Red and Near-InfraRed (NIR) light reflected by vegetation optical bands derived from Landsat 8 atmospherically corrected surface reflectance (Eq. 1).

$$NDVI = \frac{NIR-Red}{NIR+Red} \quad (1)$$

Second, the Split Window algorithm (SW)[23] is applied to retrieve LST from Landsat 8 thermal bands. The thermal bands of Landsat 8 are the bands 10 and 11, also known as TIR 1 (10.60 to 11.19µm) and TIR 2 (11.50 to 12.51µm) channels, respectively. The algorithm is a generalization of the Split-Window algorithm developed for the MODerate Resolution Imaging Spectrometer (MODIS) data[24]. The adapted version for Landsat uses a practical mathematical formula[25] (Eq. 2).

$$LST = b_0 + \left(b_1 + b_2\frac{1-LSE_{mean}}{LSE_{mean}} + b_3\frac{\Delta LSE}{LSE^2}\right)\frac{T_{ToA\,10}+T_{ToA\,11}}{2} + \left(b_4 + b_5\frac{1-LSE_{mean}}{LSE_{mean}} + b_6\frac{\Delta LSE}{LSE^2}\right)\frac{T_{ToA\,10}-T_{ToA\,11}}{2} + b_7(T_{ToA\,10} - T_{ToA\,11})^2 \quad (2)$$

The b_0, b_1, …, b_7 are the algorithm coefficients and are derived from atmospheric profile dataset using simulation codes[23]. $T_{ToA\ 10}$ and $T_{ToA\ 11}$ are Top of Atmosphere brightness temperature for bands 10 and 11. The LSE_{mean} is the mean Land Surface Emissivity of the two bands 10 and 11 (Eq. 3). ΔLSE is the difference in LSE of two bands 10 and 11(Eq. 4).

$$LSE_{mean} = \frac{LSE_{10}+LSE_{11}}{2} \quad (3)$$

$$\Delta LSE = LSE_{10} - LSE_{11} \quad (4)$$

To calculate the T_{ToA}, we first converted raw bands 10 and 11 to spectral radiance values $L_{\gamma,\,ToA}$ (Eq. 5). This later is used in the inverted Planck's radiation equation to get the resulting T_{ToA} in Celsius (Eq. 6).

$$L_{\gamma,ToA} = M_L.Q_{cal} + A_L \quad (5)$$

where M_L is the radiance multiplicative scaling factor for the given band, A_L is the radiance additive scaling factor for the given band and Qcal is raw value of the given band also known as Digital Number (DN) value.

$$T_{ToA} = \frac{K_2}{\ln(\frac{K_1}{L_{\gamma,ToA}}+1)} - 273.15 \quad (6)$$

where K_1 and K_2 are the thermal conversion constants for the given band.
To calculate *LSE*, we used equations 7 and 8 which require, in turn, the estimation of the Proportion of vegetation (*Pv*) (Eq. 9).

$$LSE_{10} = 0.0015 \times Pv + 0.9843 \quad (7)$$

$$LSE_{11} = 0.0011 \times Pv + 0.9885 \quad (8)$$

$$P_v = \frac{NDVI - NDVI_{min}}{NDVI_{max} - NDVI_{min}} \quad (9)$$

where $NDVI_{max} = 0.5$, and $NDVI_{min} = 0.2$.

b. Using raw satellite data and climate data

This approach adopts an end-end mechanism, consisting of using directly the 12 available raw Landsat 8 bands combined with climate data (T_a and R_g).

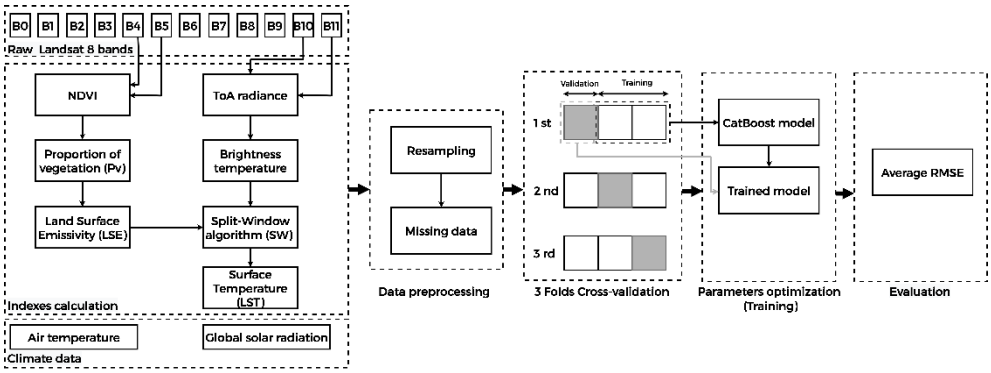

Fig. 2. The flowchart of the approach.

To evaluate the CatBoost model, the Root Mean Squared Error (RMSE) (Eq. 10) was used.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \quad (10)$$

The flowchart steps were implemented using Python language and DST[26] and ClimateFiller libraries[27].

4 Results and Discussion

Following a 3-fold cross-validation technique, the CatBoost model yielded an RMSE of 27.54 W.m^{-2} using the first approach, and 27.05 W.m^{-2} using the second approach, as indicated in Table 2.

Table 1. The RMSE values for the two data combination approaches.

Fold	Raw bands	NDVI & LST
1	25.789654	32.004610
2	28.713678	26.063326
3	26.632358	24.541034
Average	27.045230	27.536323

The results show no significant difference in terms of RMSE for both feature sets. These findings suggest that you can achieve accurate LE estimation relying solely on the raw spectral bands. This approach presents various advantages. It is computationally efficient since it eliminates the need to acquire additional data such as LST or deriving vegetation indices that may involve additional computation power. It is also practical in operational settings, especially for real-time monitoring applications where efficiency and timeliness are critical.

However, it's essential to consider whether there are specific scenarios or conditions where using indices and derived parameters might provide an advantage. Further investigation includes analysing the approach when using different land cover types (different crops) or challenging environmental conditions and testing it for other applications such as yield estimation and soil moisture retrieval.

5 Conclusion

This study examined two data-driven methods to estimate Latent heat flux (LE). The first incorporates climate data, NDVI, and LST, while the second uses climate data along with raw Landsat 8 bands. We gathered in-situ data from a flux station within our study area which is a rainfed wheat field. The data includes air temperatures, global solar radiation, and measured LE from 2015 to 2018. The CatBoost model achieved an average RMSE of 27.54 $W.m^{-2}$ for the first method and 27.05 $W.m^{-2}$ for the second, through 3-fold cross-validation. Our future scope involves generalizing the approach and testing it for other applications.

Acknowledgment

This study was supported by and conducted within the Center for Remote Sensing Applications (CRSA) (https://crsa.um6p.ma), at the Mohammed VI Polytechnic University (UM6P) in Morocco. Additionally, we would like to thank the Moroccan Ministry of Higher Education, Scientific Research and Innovation and the OCP Foundation who funded this work through the APRD research program (GEANTech). And also, we would like to recognize the contributions of the AGREEMed, PRIMA-S2-ALTOS-2018 and ASSIWAT projects in directing the outcomes of this research.

References

[1] S. Belaqziz *et al.*, “Irrigation scheduling of a classical gravity network based on the Covariance Matrix Adaptation - Evolutionary Strategy algorithm,” *Comput. Electron. Agric.*, vol. 102, pp. 64–72, Mar. 2014, doi: 10.1016/j.compag.2014.01.006.

[2] S. L. Davis, M. D. Dukes, and G. L. Miller, “Landscape irrigation by evapotranspiration-based irrigation controllers under dry conditions in Southwest Florida,” *Agric. Water Manag.*, vol. 96, no. 12, pp. 1828–1836, Dec. 2009, doi: 10.1016/J.AGWAT.2009.08.005.

[3] A. Diarra *et al.*, “Performance of the two-source energy budget (TSEB) model for the monitoring of evapotranspiration over irrigated annual crops in North Africa,” *Agric. Water Manag.*, vol. 193, pp. 71–88, Nov. 2017, doi: 10.1016/J.AGWAT.2017.08.007.

[4] S. Amani and H. Shafizadeh-Moghadam, “A review of machine learning models and influential factors for estimating evapotranspiration using remote sensing and ground-based data,” *Agric. Water Manag.*, vol. 284, p. 108324, Jun. 2023, doi: 10.1016/J.AGWAT.2023.108324.

[5] A. Subedi and J. L. Chávez, “Crop Evapotranspiration (ET) Estimation Models: A Review and Discussion of the Applicability and Limitations of ET Methods,” *J. Agric. Sci.*, vol. 7, no. 6, p. p50, May 2015, doi: 10.5539/JAS.V7N6P50.

[6] B. Duchemin *et al.*, “Agrometerological study of semi-arid areas: an experiment for analysing the potential of time series of FORMOSAT-2 images (Tensift-Marrakech plain),” *Int. J. Remote Sens.*, vol. 29, no. 17–18, pp. 5291–5299, 2008, doi: 10.1080/01431160802036482.

[7] C. Carter and S. Liang, “Evaluation of ten machine learning methods for estimating terrestrial evapotranspiration from remote sensing,” *Int. J. Appl. Earth Obs. Geoinf.*, vol. 78, pp. 86–92, Jun. 2019,

doi: 10.1016/J.JAG.2019.01.020.

[8] V. Douna, V. Barraza, F. Grings, A. Huete, N. Restrepo-Coupe, and J. Beringer, "Towards a remote sensing data based evapotranspiration estimation in Northern Australia using a simple random forest approach," *J. Arid Environ.*, vol. 191, p. 104513, Aug. 2021, doi: 10.1016/J.JARIDENV.2021.104513.

[9] Y. Liu *et al.*, "A framework for actual evapotranspiration assessment and projection based on meteorological, vegetation and hydrological remote sensing products," *Remote Sens.*, vol. 13, no. 18, p. 3643, Sep. 2021, doi: 10.3390/RS13183643/S1.

[10] V. G. Stefan, O. Merlin, S. Er-Raki, M. J. Escorihuela, and S. Khabba, "Consistency between In Situ, Model-Derived and High-Resolution-Image-Based Soil Temperature Endmembers: Towards a Robust Data-Based Model for Multi-Resolution Monitoring of Crop Evapotranspiration," *Remote Sens. 2015, Vol. 7, Pages 10444-10479*, vol. 7, no. 8, pp. 10444–10479, Aug. 2015, doi: 10.3390/RS70810444.

[11] Z. Chen, R. Shi, and S. Zhang, "An artificial neural network approach to estimate evapotranspiration from remote sensing and AmeriFlux data," *Front. Earth Sci.*, vol. 7, no. 1, pp. 103–111, Mar. 2013, doi: 10.1007/S11707-012-0346-7/METRICS.

[12] L. A. Reyes Rojas, I. Moletto-Lobos, F. Corradini, C. Mattar, R. Fuster, and C. Escobar-Avaria, "Determining actual evapotranspiration based on machine learning and sinusoidal approaches applied to thermal high-resolution remote sensing imagery in a semi-arid ecosystem," *Remote Sens.*, vol. 13, no. 20, p. 4105, Oct. 2021, doi: 10.3390/RS13204105/S1.

[13] P. S. Käfer *et al.*, "Artificial neural networks model based on remote sensing to retrieve evapotranspiration over the Brazilian Pampa," *https://doi.org/10.1117/1.JRS.14.038504*, vol. 14, no. 3, p. 038504, Sep. 2020, doi: 10.1117/1.JRS.14.038504.

[14] P. Hao, L. Di, E. Yu, L. Guo, Z. Sun, and H. Zhao, "Using machine learning and trapezoidal model to derive All-weather et from Remote sensing Images and Meteorological Data," *2021 9th Int. Conf. Agro-Geoinformatics, Agro-Geoinformatics 2021*, Jul. 2021, doi: 10.1109/AGRO-GEOINFORMATICS50104.2021.9530341.

[15] Y. Liu, S. Zhang, J. Zhang, L. Tang, and Y. Bai, "Assessment and Comparison of Six Machine Learning Models in Estimating Evapotranspiration over Croplands Using Remote Sensing and Meteorological Factors," *Remote Sens. 2021, Vol. 13, Page 3838*, vol. 13, no. 19, p. 3838, Sep. 2021, doi: 10.3390/RS13193838.

[16] S. Ha and H. Jeong, "Unraveling hidden interactions in complex systems with deep learning," *Sci. Reports 2021 111*, vol. 11, no. 1, pp. 1–13, Jun. 2021, doi: 10.1038/s41598-021-91878-w.

[17] R. G. Allen, L. S. Pereira, D. Raes, and M. Smith, "Crop Evapotranspiration-Guidelines for Computing Crop Water Requirements. FAO Irrigation and Drainage Paper 56. Food and Agriculture Organization of the United Nations (FAO).," Rome, 1998.

[18] B. Ait Hssaine, O. Merlin, J. Ezzahar, N. Ojha, S. Er-Raki, and S. Khabba, "An evapotranspiration model self-calibrated from remotely sensed surface soil moisture, land surface temperature and vegetation cover fraction: Application to disaggregated SMOS and MODIS data," *Hydrol. Earth Syst. Sci.*, vol. 24, no. 4, pp. 1781–1803, Apr. 2020, doi: 10.5194/HESS-24-1781-2020.

[19] T. E. Twine *et al.*, "Correcting eddy-covariance flux underestimates over a grassland," *Agric. For. Meteorol.*, vol. 103, no. 3, pp. 279–300, Jun. 2000, doi: 10.1016/S0168-1923(00)00123-4.

[20] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Curran Associates Inc., 2018, pp. 6639–6649. doi: 10.5555/3327757.3327770.

[21] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-Augu, pp. 785–794, Aug. 2016, doi: 10.1145/2939672.2939785.

[22] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Curran Associates Inc., 2017, pp. 3149–3157. doi: 10.5555/3294996.3295074.

[23] J. C. Jimenez-Munoz, J. A. Sobrino, D. Skokovic, C. Mattar, and J. Cristobal, "Land surface temperature retrieval methods from landsat-8 thermal infrared sensor data," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 10, pp. 1840–1843, 2014, doi: 10.1109/LGRS.2014.2312032.

[24] Z. Wan, "New refinements and validation of the collection-6 MODIS land-surface temperature/emissivity product," *Remote Sens. Environ.*, vol. 140, pp. 36–45, Jan. 2014, doi: 10.1016/J.RSE.2013.08.027.

[25] M. Q. U. Sajib and T. Wang, "Estimation of Land Surface Temperature in an Agricultural Region of Bangladesh from Landsat 8: Intercomparison of Four Algorithms," *Sensors (Basel).*, vol. 20, no. 6, Mar. 2020, doi: 10.3390/S20061778.

[26] C. El Hachimi, S. Belaqziz, S. Khabba, and A. Chehbouni, "Data Science Toolkit: An all-in-one python library to help researchers and practitioners in implementing data science-related algorithms with less effort," *Softw. Impacts*, vol. 12, p. 100240, May 2022, doi: 10.1016/J.SIMPA.2022.100240.

[27] C. El Hachimi *et al.*, "ClimateFiller: A Python framework for climate time series gap-filling and diagnosis based on artificial intelligence and multi-source reanalysis data," *Softw. Impacts*, vol. 18, p. 100575, Nov. 2023, doi: 10.1016/j.simpa.2023.100575.