data

MDPI

Article

A Data Storage, Analysis, and Project Administration Engine (TMFdw) for Small- to Medium-Size Interdisciplinary Ecological Research Programs with Full Raster Data Capabilities

Paulina Grigusova [1,*], Christian Beilschmidt [2], Maik Dobbermann [1], Johannes Drönner [2], Michael Mattig [2], Pablo Sanchez [3], Nina Farwig [4] and Jörg Bendix [1]

[1] Laboratory for Climatology and Remote Sensing (LCRS), Department of Geography, University of Marburg, D-35032 Marburg, Germany; bendix@staff.uni-marburg.de (J.B.)

[2] Geo Engine GmbH, Am Kornacker 68, D-35041 Marburg, Germany; christian.beilschmidt@geoengine.de (C.B.); mattig@mathematik.uni-marburg.de (M.M.)

[3] Instituto Nacional de Biodiversidad (INABIO), Pje. Rumipamba N. 341 y Av. de los Shyris (Parque La Carolina), Quito 170102, Ecuador; pablo.sanchez@biodiversidad.gob.ec

[4] Department of Biology, Conservation Ecology, University of Marburg, D-35032 Marburg, Germany; farwig@staff.uni-marburg.de

* Correspondence: paulina.grigusova@staff.uni-marburg.de

Abstract: Over almost 20 years, a data storage, analysis, and project administration engine (TMFdw) has been continuously developed in a series of several consecutive interdisciplinary research projects on functional biodiversity of the southern Andes of Ecuador. Starting as a "working database", the system now includes program management modules and literature databases, which are all accessible via a web interface. Originally designed to manage data in the ecological Research Unit 816 (SE Ecuador), the open software is now being used in several other environmental research programs, demonstrating its broad applicability. While the system was mainly developed for abiotic and biotic tabular data in the beginning, the new research program demands full capabilities to work with area-wide and high-resolution big models and remote sensing raster data. Thus, a raster engine was recently implemented based on the Geo Engine technology. The great variety of pre-implemented desktop GIS-like analysis options for raster point and vector data is an important incentive for researchers to use the system. A second incentive is to implement use cases prioritized by the researchers. As an example, we present machine learning models to generate high-resolution (30 m) microclimate raster layers for the study area in different temporal aggregation levels for the most important variables of air temperature, humidity, precipitation, and solar radiation. The models implemented as use cases outperform similar models developed in other research programs.

Dataset: The link to the datasets is as follows: https://respect.app.geoengine.io (accessed on 2 December 2024).

Keywords: working database, big raster data, raster engine, use case, area-wide microclimate

Citation: Grigusova, P.; Beilschmidt, C.; Dobbermann, M.; Drönner, J.; Mattig, M.; Sanchez, P.; Farwig, N.; Bendix, J. A Data Storage, Analysis, and Project Administration Engine (TMFdw) for Small- to Medium-Size Interdisciplinary Ecological Research Programs with Full Raster Data Capabilities. *Data* **2024**, *9*, 143. https://doi.org/10.3390/data9120143

Academic Editor: Juanle Wang

Received: 1 October 2024
Revised: 26 November 2024
Accepted: 30 November 2024
Published: 6 December 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Research data should be preserved and curated for future reuse in a sustainable manner, for example, following the FAIR (Findable, Accessible, Interoperable, Reusable) principles [1] or other models [2]. In comparison to FAIR, the Linked Open Data 5-star paradigm strongly supports open data and does not make data reuse dependent on licensing agreements [3]. This is important because open and reproducible science requires free access to research data [4,5]. The need for open data in all areas of science is becoming increasingly urgent, especially with the growing demand for data-driven applications [6]. This all holds for interdisciplinary biodiversity and other ecological research programs [7–9]. On the global to national scale, several repositories for long-term storage of biodiversity data are meanwhile established, such as the Global Biodiversity Information Facility (GBIF) [10], to name just one of the most prominent data infrastructures. In Germany, for instance, a multi-cloud platform based on the German Federation for Biological Data (GFBio) [11] project, the NFDI4Biodiversity (NFDI = German National Research Data Infrastructure), is currently developed in order to provide a long-term repository of data for biodiversity and ecology research, with a main aim to mobilize national data from research and collections [12,13]. However, challenges are becoming larger in the age of big data and the Internet of Things where not only species and other tabular data are of interest but a variety of big data sources (genomics, remote sensing, numerical models, audio and video, etc.) [14] have to be managed and offered in such a way that users with usually restricted data literacy can provide and use data in a differentiated and uncomplicated manner [15–21].

Beyond such global and national data repositories, data collection, analysis, and provision often start in longer-term interdisciplinary research programs where proper data management and stewardship are a major task of project management [22,23]. One example in German interdisciplinary biodiversity research is the BExIS system [24,25], which is an open-source research data management system to support interdisciplinary research projects with multiple subprojects following the FAIR principle. It has been successfully implemented for major German programs such as the Biodiversity Exploratories, the Jena Experiment, and the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig [25]. Another widely used system is the TMF Data Warehouse (TMFdw), which was developed in another longer-term biodiversity-related research program [24] that started in 1997 in the biodiversity hotspot of the SE Ecuadorian Andes. In this program, data types and volumes have changed as the focus of the research program has changed, and so have the data management requirements. Research started in 1997 with an abiotic and biotic inventory phase of the mega-diverse mountain rainforest ecosystem [26] and then turned to the analysis of ecosystem functioning [27]. Consecutive activities focused on ecosystem services under environmental change by including more and more area-wide datasets [28,29]. This led to the development of knowledge transfer solutions for sustainable land use as well as area-wide biodiversity and atmospheric monitoring [30–33]. Meanwhile, a follow-up program (Research Unit RESPECT; environmental changes in biodiversity hotspot ecosystems of South Ecuador: RESPonse and feedback effECTs [34]) investigates future environmental changes in the biodiversity hotspot by means of data-driven integrated Response–Effect Framework analysis and a new generation of biodiversity-informed land surface models [35]. After more than 20 years of research, around 14 GB of tabular data have been collected in the TMFdw. Particularly, the running phase is now challenged by the integration of big geospatial data and the need for data enrichment and easy-to-use services by user-defined use cases. The TMFdw data store is designed to provide a flexible and standardized platform for managing a wide range of ecological datasets and is aimed at supporting researchers, environmental analysts, and policymakers. The primary aim is to facilitate interdisciplinary ecological research by storing, integrating, and providing access to various types of data, including biological data (such as biodiversity measurements and vegetation information) and environmental data (such as climate observations and topographic information).

The main objective of the paper is (1) to document the development steps of the TMFdw and (2) present the new features related to the management of large raster datasets as well as (3) the recent integration of a central user-driven use case, namely, the generation of high-resolution climate raster data from point-based station measurements with machine learning.

The current paper is structured as follows. First, information about the history of development and successful implementations of the TMFdw is given; then, the new tool to integrate big area-wide raster data using the Geo Engine [33] is introduced. The paper will close with a description of the use case prioritized by the researchers to offer area-wide microclimate data.

2. The Data Warehouse TMFdw

2.1. Development and Implementation

The need to provide data access for the interdisciplinary program in SE Ecuador resulted in the development of the first meta-database connected to a simple file system in the year 2001 [36]. The increasing diversity of data and the need to make it available to the research program and the public led to the development of the TMFdw in 2007, which has continued to evolve ever since (see Figure 1 [24]).

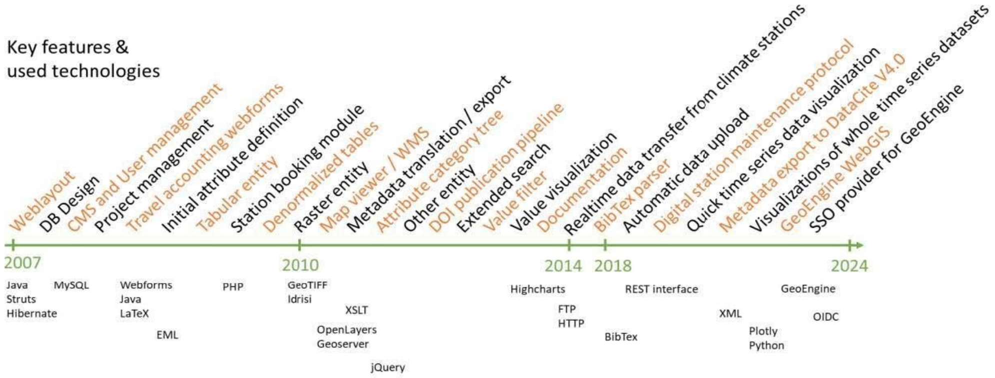

Figure 1. Development phases of the TMFdw.

The TMFdw technology will be freely available through the GitLab page [37] of the developing research group. Due to the different requirements of the projects and to meet new demands such as automatic uploading, better data visualization, and analysis tools (Figure 1) have been continuously developed. As a result, the TMFdw is meanwhile capable of hosting a diverse set of data from different scientific disciplines such as biodiversity, ecology, hydrology, atmospheric sciences, soil sciences, and health research, and thus, it was successfully applied in different collaborative programs in different fields of environmental research (Table S1).

The main component of the TMFdw is a web-based system built using enterprise-level Java technology. It operates on a Tomcat server, stores information in a MySQL (My Structured Query Language) database, and uses the Hibernate framework for efficient data handling. The platform includes a REST API (representational state transfer), enabling users to upload data to ongoing projects easily. Recent updates have added support for large-scale raster data through integration with the Geo Engine. Additionally, a Single Sign-On feature using OpenID Connect has been implemented, allowing seamless user access across the system. Uploaded data, including raw files, processed information, and metadata, are regularly backed up on secure tape storage managed by the university's IT department.

The data warehouse development follows the standards of the FAIR principle [1] that emphasizes the importance of data reuse. (i) Data should be findable. The TMFdw is a certified DOI (Digital Object Identifier) data center. The datasets and their metadata are provided via the website of the data warehouse, indexed by various web crawlers, and datasets that have a self-registered DOI that are findable in DOI catalogs. The data warehouse for the SE Ecuadorian Andes itself is registered in the archive of data repositories re3data.org [38,39]. A DOI data center guarantees access to data for at least 10 years. For longer-term storage, we will provide data migration to a long-term archive, preferably

NFDI4Biodiversity. (ii) Data should be accessible and reusable. Datasets will be made publicly available (converted to open data) by the decision of the member assembly after the end of a funding period (3–4 years) of a fixed-term phase of the research program. Before, interested researchers can access the data owner (provided with the metadata) with direct permission to reuse the data. Data use and reuse are regulated by a data use agreement [40], which must be accepted by data users before downloading the datasets. Interested users can access open data via the data warehouse with a simple human verification via email. (iii) Data should be Interoperable. We save our metadata in the EML (Ecological Metadata Language) [41,42], which is openly available and is based on and compatible with the Darwin Core [43] format.

The data flow in the TMFdw is depicted in Figures 2 and 3. In general, unstandardized and heterogeneous data with different data acquisition methods through instruments or manual collection must be transformed through conversion and manual or automatic preprocessing to standardized and homogeneous datasets in the data warehouse. First, standardized EML metadata need to be generated by the data provider as follows: Who has collected What data When and Where, and Which methods were used. The interface provides a general block of metadata, which includes the Who, When, and Where, and a section of metadata for each attribute (column) that should contain information about measurement methods and instrumentation. After that, data can be uploaded in CSV format for tabular data, GeoTiff is used for raster data and shape files are used for vector data. The user has to only create a dataset for this type of data once manually in the data warehouse and can set up an upload script to attach new data to this dataset via this interface and the unique resource ID. Of course, the attributes of the new data have to match the ones in the existing dataset. For logger data, e.g., from our automatic weather stations (AWS) in the study area that need to be uploaded continuously, we implemented an automated uploading possibility via a REST interface and an automatic quality control [44].

At the moment, the TMFdw holds around 14 GB of tabular data. To categorize the data ontologically, we provide over 1000 categories grouped in 100 category nodes. Data providers can select from over 1300 attributes to map to their data with the possibility to create new attributes where needed. Over the years, we have given numerous data workshops to train new researchers. Users have two options for analyzing the data. They can use the analysis and visualization functions provided by the TMFdw and its raster engine via the user interface. If a non-implemented analysis is required, they can interact with the data through an online user interface or using the GeoEngine package in Python, allowing them to perform analysis and visualization directly within the system. Additionally, users can download the data to use in their own software tools through well-defined interfaces. The available functionality includes operators for aggregation (space and time), raster–vector combination, data filters, and raster and vector calculators; see docs.geoengine.io (accessed on 2 December 2024) for more details. The data stored in the repository are generally open source, providing broad access to researchers, stakeholders, and the public. However, certain datasets may be subject to a limited embargo period, particularly if they are unpublished data needed for a PhD thesis. This provides some protection for PhD students to publish their data first. During this time, however, access to individual datasets is possible by contacting the data owner who is identified in the metadata. After this embargo period, the data are set to open access.

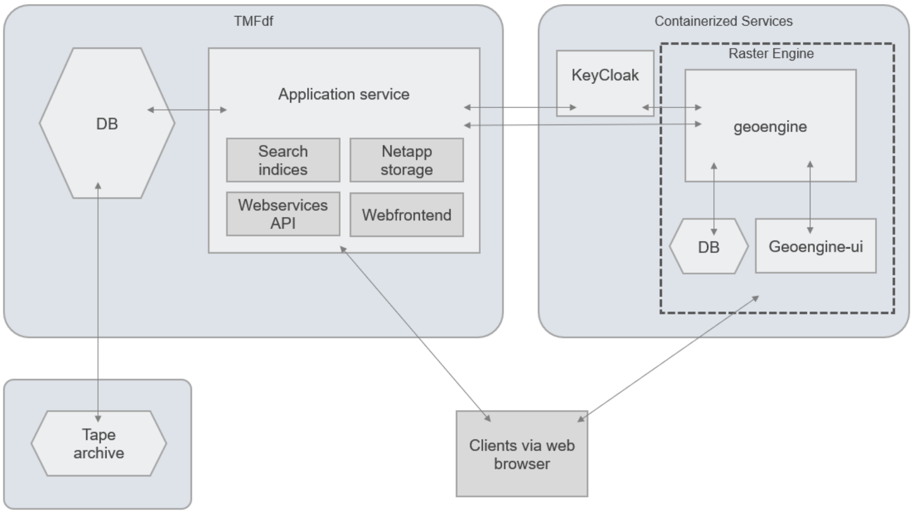

Figure 2. System architecture diagram. DB (database) (described in detail in Section 2.2.1).

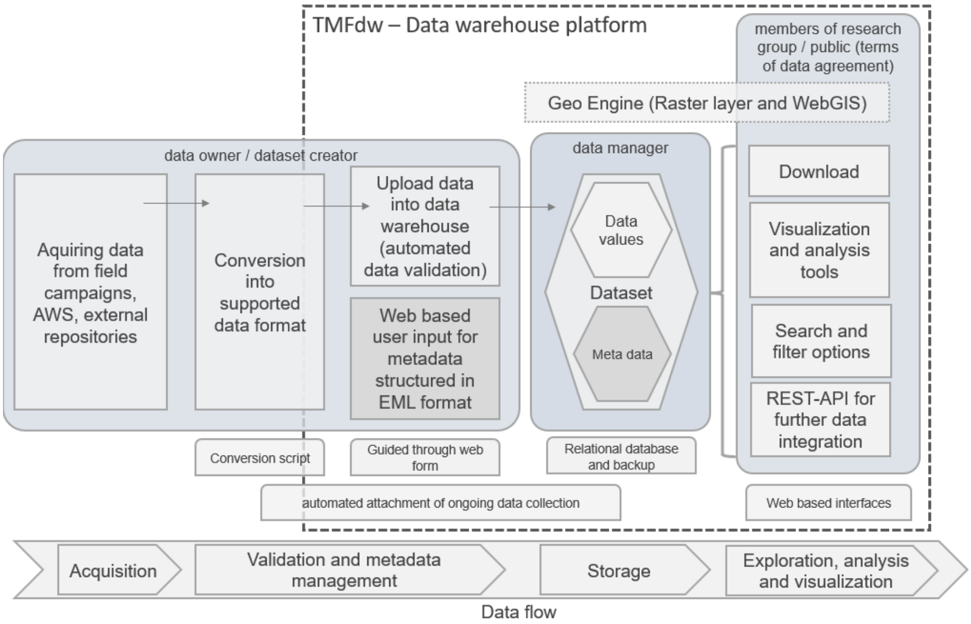
TMFdw – Data warehouse platform
members of research group / public (terms of data agreement)
Geo Engine (Raster layer and WebGIS)
data owner / dataset creator
Aquiring data from field campaigns, AWS, external repositories
Conversion into supported data format
Upload data into data warehouse (automated data validation)
Web based user input for metadata structured in EML format
data manager
Data values
Dataset
Meta data
Download
Visualization and analysis tools
Search and filter options
REST-API for further data integration
Conversion script
Guided through web form
Relational database and backup
automated attachment of ongoing data collection
Web based interfaces
Acquisition
Validation and metadata management
Storage
Exploration, analysis and visualization
Data flow

Figure 3. Data warehouse platform. AWS: automatic weather station (described in detail in Section 2.2.1).

2.2. The Novel TMFdw Raster Engine Using Geo Engine Technology

In the current program, a new challenge was to include big, raster data [34], which was not possible with the recent structure of the TMFdw. As described below (Section 2.2.1), the Geo Engine greatly improves the very basic raster data handling of the existing TMFdw implementation, providing web GIS functionality and better raster data storage capabilities. Relevant raster data are (i) customized climate and land-use model data for testing and forcing of a newly developed land surface model (LSM) HUMBOL-TD (Hydroatmo Unified Model of Biotic interactions Organic Matter and Local Trait Diversity), [45], (ii) high-resolution area-wide data from HUMBOL-TD model runs, and (iii) the continuously AI-generated high-resolution microclimate raster data using the automatically harvested tabular AWS data from the implemented use case For the model applications, we strive for a grid resolution of 1 km; for further analyses (land use changes, area-wide microclimatic data), we work on a 30 m grid resolution. Depending on the application, the temporal resolution of datasets ranges from 5 min to annual aggregation. To avoid unnecessarily high development costs, we decided to include and adapt a readily available framework for storing and analyzing big raster data for the ecological domain, the Geo Engine [46].

2.2.1. Using the Geo Engine as the TMFdw Raster Engine

The Geo Engine features a data provider concept that allows it to connect to remote raster and vector data as well as local datasets. The TMFdw raster engine provides common locally stored vector datasets, like project areas and plots and regional political boundaries, as well as large raster data, e.g., climate model and satellite images, accessed and downloaded via APIs from the cloud. It also provides a means for project members to share their results as new datasets within the TMFdw raster engine with other members. This allows project results to be easily incorporated into new analyses. In addition to data, the system offers workflow processing capabilities and interfaces for visualization and Python programming environments. The Geo Engine is an open-source platform for working with remote sensing and model raster data and for combining or integrating different types of spatial data (such as raster, vector, and tabular data) into a unified analysis or visualization. The idea is to alleviate the purely technical parts of raster data problems and processing, such as big data processing and time series data handling. In general, the Geo Engine connects to different data sources and provides harmonized access to them. The platform allows the definition of data pipelines using workflows on such harmonized geo time series.

The main features of the Geo Engine are summarized in Figure 4. The Geo Engine has several key features designed to handle large, complex datasets efficiently. Instead of working with single images or files, it focuses on processing time series data, such as monthly images or other datasets that change over time. This approach is useful because it allows the engine to automatically manage updates and perform calculations over time, making it easier for users to work with evolving data. The engine uses a special way of processing data called chunk-wise processing. This means it breaks down large datasets into smaller pieces and handles each piece separately. For example, if you have a time series of satellite images, the engine processes them step by step, band by band, and tile by tile. This method is also applied to vector data (like points or shapes on a map), dividing them into chunks based on their size or complexity. Users can apply various tools and calculations, such as filters, heatmaps, or combining data from different sources. For instance, a user might load satellite images and compute vegetation indexes, like the NDVI, and combine these results with map features, such as fields or regions, to create a dataset ready for modeling. The Geo Engine does not store processed layers permanently on disk. Instead, it generates them on demand, which is why they are called virtual layers. This allows users to access the latest data without needing to handle updates manually. Users can access data through standard web-based mapping tools or programming environments, like Jupyter Notebooks. The system also connects easily with external data sources, allowing users to pull in data from other services without needing to store everything themselves. The Geo Engine's web interface includes features like maps, data tables, and tools for creating custom workflows. It is designed to be user friendly, especially for those familiar with any desktop GIS software. Users can browse the data catalog, apply various tools, see the results on a map, and either download the processed data or continue working with it using Python 3 in Jupyter Notebooks 7.2.2. The virtual layer feature ensures that the data provided is always up to date, for example, by automatically updating daily temperature readings for a specific area. This means users obtain ready-to-use data without needing to handle complex processing themselves.

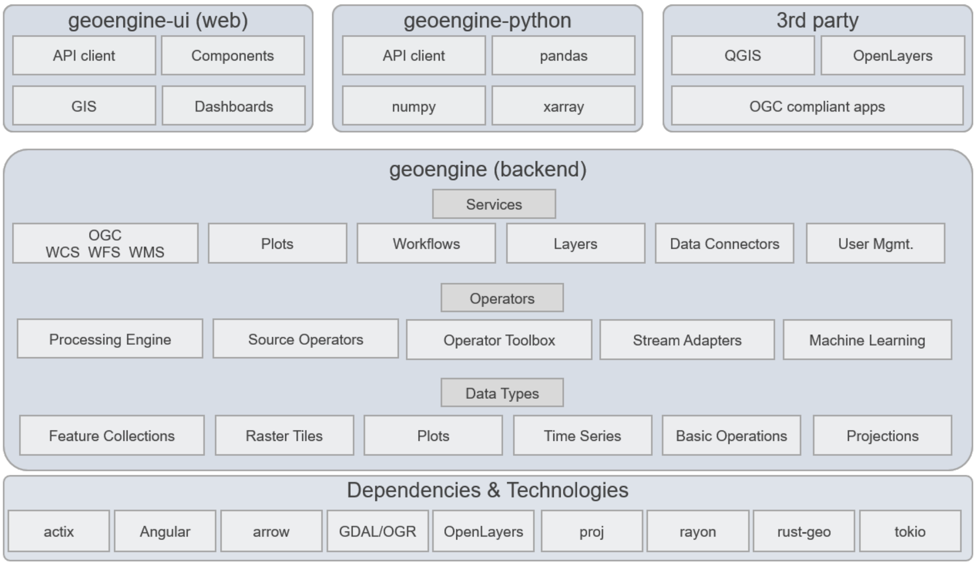

Figure 4. Component diagram showing the Geo Engine and its component blocks. geoengine-UI (web interface) (described in detail in Section 2.2.2).

2.2.2. Deployment and Connection of the Geo Engine as the TMFdw Raster Engine

The aim was to smoothly connect the raster processing tool with the main data system, making it easy for users to work with spatial data, similar to using desktop mapping software. The first step was creating a consistent user interface, matching the look and feel of the existing system. The second step was setting up a common login system so users do not need to sign in multiple times. The platform is built with different components working together. As a cloud-ready system, the Geo Engine makes use of open Open ID Connect [47] containers to deploy the processing backend, databases, user interface, and KeyCloak [48].

KeyCloak is an open-source identity and access management solution for applications and services, providing features like single sign-on (SSO), user federation, and social login. This is deployable for scalability via three tools. The first tool is Docker, a platform for developing, shipping, and running containerized applications, enabling consistent environments across different systems. The second is Podman, a daemonless, open-source container engine similar to Docker, offering container management with enhanced security features and compatibility with Kubernetes. The third is Kubernetes, which is an open-source container orchestration platform for automating the deployment, scaling, and management of containerized applications.

When a user logs in, the Geo Engine automatically creates a linked user profile in the background. This way, users can move between the main data system and the raster engine seamlessly. The system remembers their work, such as recent data layers or calculations. Users can also easily access their data in programming tools, like Jupyter Notebooks, using an access token. If a user's account is deleted from the main system, they can no longer access the connected Geo Engine tools. This setup ensures a smooth experience while keeping data secure and easily accessible across the platform.

The look and feel of the raster engine are also important to be considered as a joint service offering. Thus, Geo Engine's web applications toolkit was used to define common elements, such as the RESPECT logo, and common colors, such as RESPECT's project main colors, which are also used in the other parts of the TMFdw. This can be seen in Figure 5. Additionally, a floating button on top always links back to the homepage of the TMFdw. Alongside the SSO, this builds a seamless user experience.

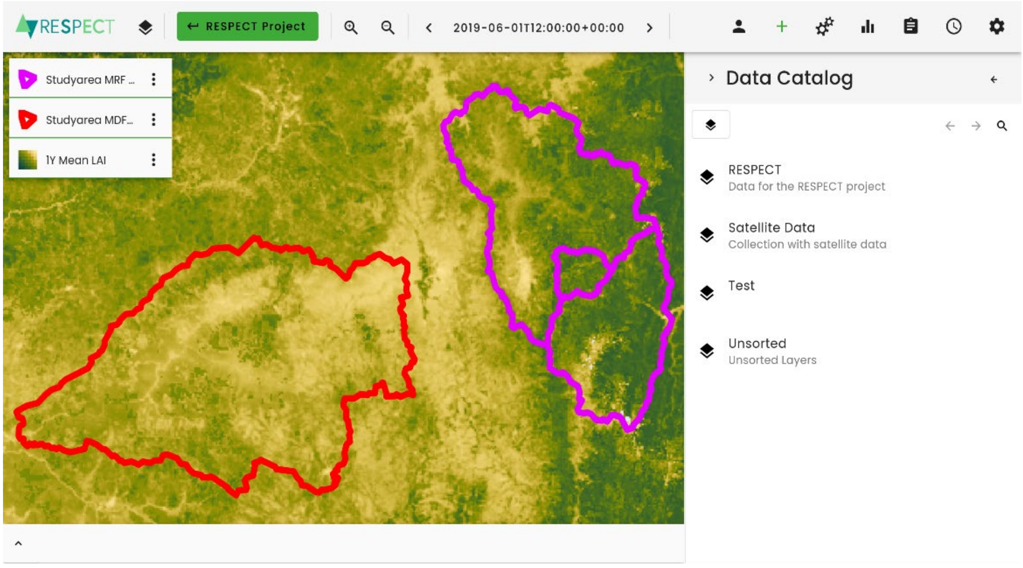

Figure 5. Graphical user interface of the raster engine.

2.2.3. The Spatial Data Catalog

One important aim of the Data Portal is to provide the project members with a single access point to commonly used remote sensing and project data. Table S2 lists third-party data already included to force and test the HUMBOL-TD LSM and describes them in further detail. The integration in the raster engine was conducted in three ways: manually (M), semi-automatically (SA), and automatically (A). Project-specific data, e.g., polygons of the study area, were uploaded into the portal and provided as layers to all users. Other data, e.g., Sentinel-2 multispectral imagery, have a connection to a STAC (SpatioTemporal Asset Catalog) service of Copernicus that can access the complete time series at any time. No up-front storage is required here. However, to speed up certain processes, users can store parts of the time series as a new, derived dataset. Data from ECMWF (European Centre for Medium-Range Weather Forecasts) and NASA could only be ingested semi-automatically since they do not offer a direct access service but rather transform data requests into a downloadable package, e.g., via a zip file as an e-mail notification. Thus, we chose to prefetch the years 2018–2023, after which an ingestion pipeline annotates the datasets with Geo Engine's metadata, e.g., defining the time series metadata, projections, and data types. All data are provided in the data catalog of the TMFdw raster engine. In the case of (A), new data will be continuously and automatically harvested to always make the latest version available to the users.

As an example, we can obtain Sentinel-3 LAI (Leaf Area Index) and MODIS NDVI products from the raster engine's data catalog (cf. Figure 6). We can display them on top of each other but also side by side. Figure 7 shows the workflow computing the mean LAI and mean NDVI value for the study areas polygons of the Mountain Rain Forest (MRF) and the Mountain Dry Forest (MDF) domain. On the right, two plots over the derived time series are displayed. On the top, a scatter plot of the LAI and NDVI for the domains is shown. At the bottom, the LAI values of the MDF are plotted over time. The workflow in Figure 7 shows how the raster data are combined with the polygon data.

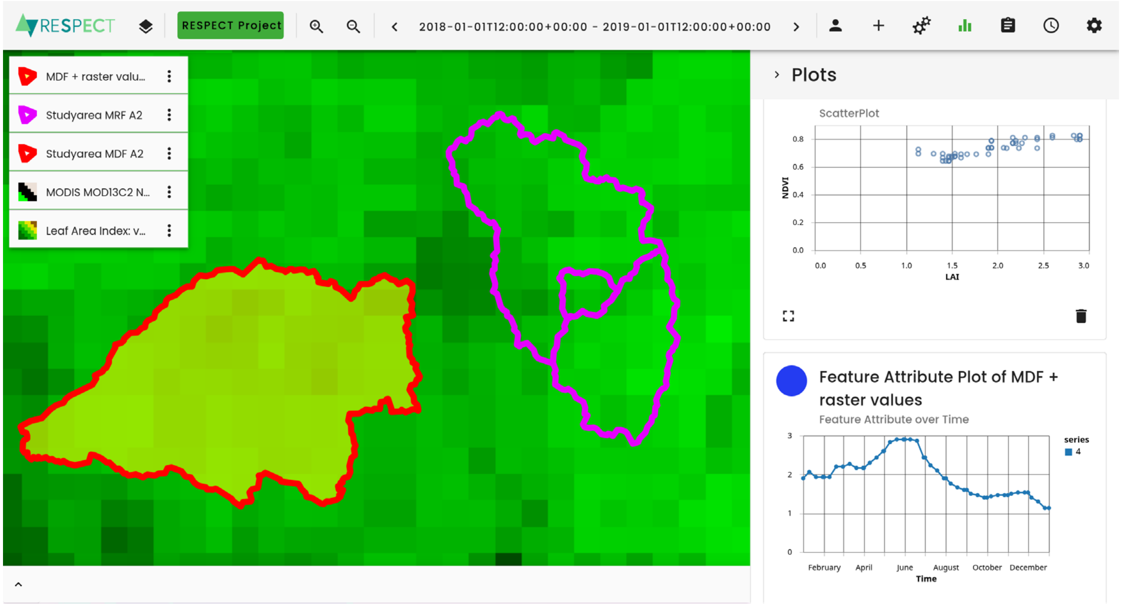

Figure 6. Map display showing NDVI and LAI values.

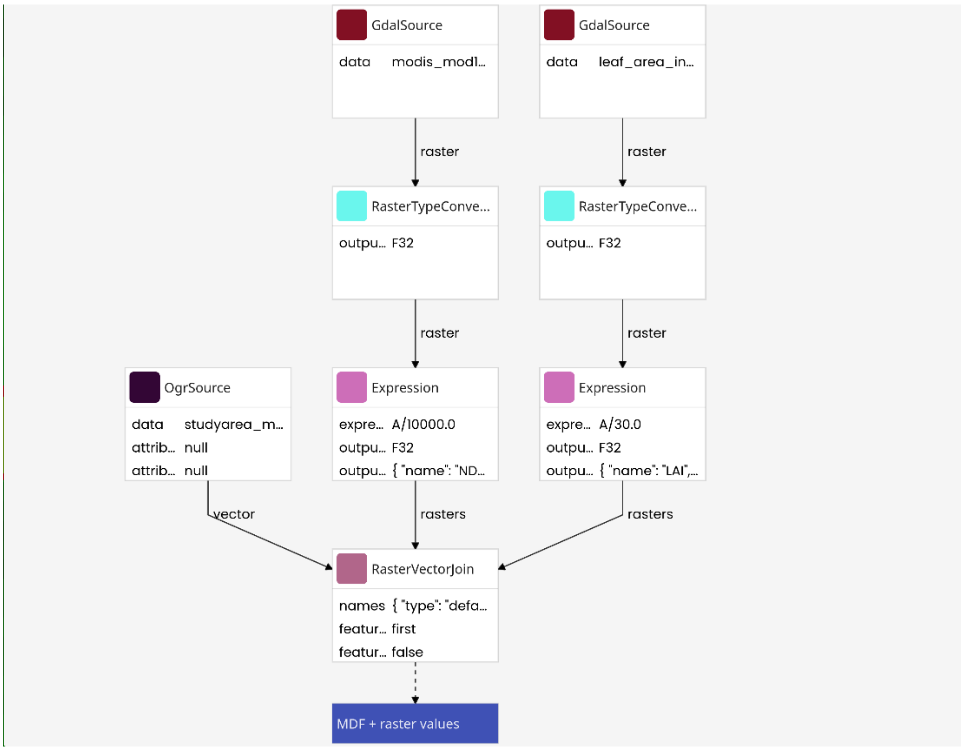

Figure 7. Examplary workflow of computing the relationbetween layers.

3. Use Case: Area-Wide Microclimate Data in High Resolution

In a survey, the members of the Research Unit RESPECT, which focuses on interdisciplinary biodiversity research programs, were asked what kind of support they expected from the new TMFdw. They unanimously favored a use case that would make it possible to generate spatially high-resolution microclimate data of the study area from the point data of the program's automatic weather stations (AWSs). These data should be used to enrich and jointly evaluate biodiversity data from any study plot that do not have their own AWS. The most important and thus target variables from an ecological point of view are air temperature, precipitation, air humidity, and solar radiation. The approach uses the tabular AWS data measured in the past and stored in the TMFdw to train and test models with spatial predictors. The models shall then be automatically applied to the continuously incoming AWS data so that the automatic upload of the AWS data from the study area results in a continuous generation of high-resolution microclimate raster data for the entire study area (Figure 8).

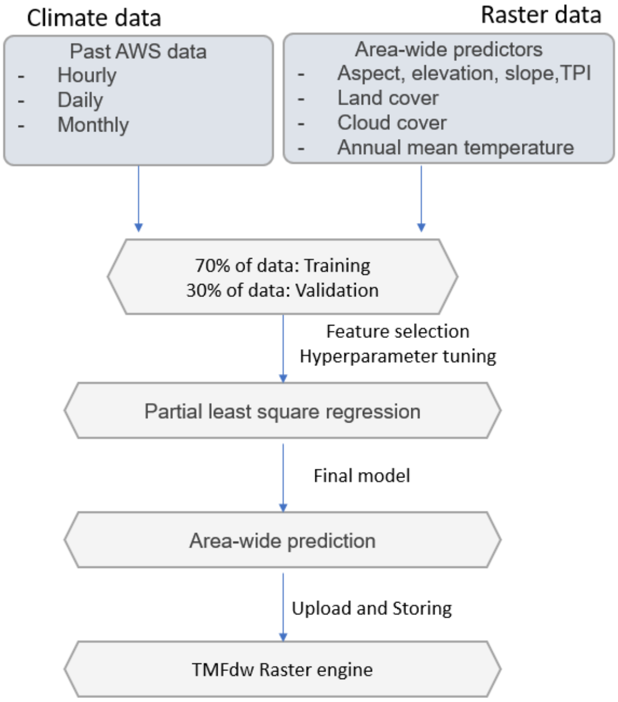

Figure 8. Schematic representation of the workflow implemented for the use case area-wide microclimate data.

3.1. Data and Method

As a basis for training, we used 11 AWSs operated for several years in the Andes of southern Ecuador (Table S3). The observation level was 2 m above ground for temperature, humidity, and radiation, and 1 m for rainfall. The PLSR models were developed for the Mountain Rain Forest domain (MRF) using AWS data between 2022 and 2023. The MRF stations represent an elevation gradient comprising natural ecosystems (forest, subparamo) and the anthropogenic replacement system (pasture).

The definition of the target variables was part of the use case survey among the scientists in the research group [49] who mainly come from the field of organismic biology. The definition was also based on previous experiences from targeted data requests to the data manager for specific ecological–statistical analyses where climate data were used as predictors (The most frequently mentioned, ecologically relevant microclimate variables here were temperature and precipitation, followed by humidity (or saturation deficit) and solar radiation (as a proxy for PAR). Additional microclimate variables were neither requested nor mentioned in the user survey. The target variables are available in 5–60 min temporal resolution at the AWS and were aggregated into hourly, daily, and monthly averages or totals. To train the PLSR model, absolute values were used for air temperature and radiation, while log transformation was applied to precipitation and humidity data. The logarithmic transformation helps to stabilize variance and approximate normality, taking into account the skewness and heteroscedasticity typical of meteorological data, thus improving the interpretability and robustness of the models.

The selected area-wide predictors are related to topography and land cover and to long-term averages of climate variables for the study area taken from the TMFdw raster engine (Table 1). The topography-related area-wide predictors rely on an Airborne Laser Scanning (ALS) campaign that yielded the digital elevation model [50]. We opted to use

Airborne Laser Scanning (ALS) data instead of publicly available DEMs due to several reasons. The ALS campaign allowed us to collect data specific to the study region with direct validation against ground control points. This localized data collection reduces potential inaccuracies associated with using global DEM products, which might not account for specific local variations or have outdated information. While it is true that ALS data might not always represent the Earth's surface accurately in densely vegetated areas, the technology's ability to penetrate vegetation and capture ground points provides a more accurate representation of the terrain than optical satellite-based DEMs, which often include canopy height as part of the elevation data. Our study is part of a larger research consortium with multiple subprojects that require a standardized approach to data. Using ALS-derived DEM data ensures consistency across all subprojects, facilitating integrated analyses and comparisons. Relying on different DEM sources could introduce variability due to differences in resolution, acquisition time, and processing methods, complicating collaborative efforts within the consortium. The slope predictor is calculated from this DEM according to [51], the aspect is derived according to [52], and the Topographic Position Index (TPI) is calculated according to [53].

The model used for this use case is Partial Least Squares Regression (PLSR). It is best at dealing with multicollinearity and small sample sizes, which are common in ecological datasets [54]. PLSR builds latent variables that capture the maximum covariance between predictors and response variables, optimizing predictive performance without overfitting. The "caret" package in R with the "pls" method as a modeling technique was used. We chose Partial Least Squares Regression (PLSR) for its ability to handle collinear, high-dimensional predictors, common in remote sensing and environmental data. PLSR is robust against noise, interpretable, and integrates well with our existing TMFdw workflows, making it ideal for extracting relevant information without overfitting [55].

Recursive feature elimination (RFE) was applied during model development to enhance performance by identifying and retaining only the most relevant predictors. RFE systematically removes less important variables and evaluates the model's performance, iterating until the optimal set of predictors is identified.

Cross-validation was employed to ensure the robustness and generalizability of the model. Specifically, we used k-fold cross-validation, where the data were partitioned into five subsets (each including two or three AWSs), with each subset being used as a test set once while the model was trained on the remaining four subsets. This approach helped to reduce overfitting and provided a more reliable estimate of the model's performance across different subsets of data. The performance of the model was evaluated using the root mean squared error (RMSE) and R-squared (R2) metrics. The metrics were estimated iteratively using the test data subsets excluded from training.

Finally, the models were intended to predict the target variables in the Mountain Rain Forest area for all retrospective and new incoming AWS data. For log-transformed variables, the values were transformed back to the original data space. Separate models were used to predict hourly, daily, and monthly data. The spatial resolution of the modeled microclimate raster data is 30 m.

Model outputs are provided via a network drive, which is connected to the Geo Engine as a data source. Scripts are then used to import or update the metadata needed to load them as a time series.

Table 1. List of area-wide predictors reused for the models of the four selected target variables (ALS = Airborne Laser Scanning campaign, X = predictor used).

Target Variable Predictors	Air Temperature	Precipitation	Humidity	Radiation
Elevation (ALS)	X	X	X	X
Slope [52]	X	X	X	X

Aspect [53]	X	X	X	X
TPI [54]	X	X	X	X
Land cover [56]	X	X	X	X
Mean annual temperature [57]	X			
Mean annual relative humidity [58]			X	
Mean annual cloud frequency [59]		X		

3.2. Quality of the Use-Case Models, Results, and Discussion

The quality of the best predictive models is generally very good based on the interpretation of the R2 and RMSE values [60], which were used to describe the goodness of fit. The RMSE varies with the target variables and the temporal aggregation (Table 2, Figure 9). In general, hourly predictions are more difficult than for higher aggregation levels and, therefore, tend to be less accurate. Air temperature reveals the highest accuracy, followed by radiation. Precipitation models perform a bit better than air humidity, except for the hourly aggregation due to the high spatio-temporal dynamics of mostly convective rain cells in the area.

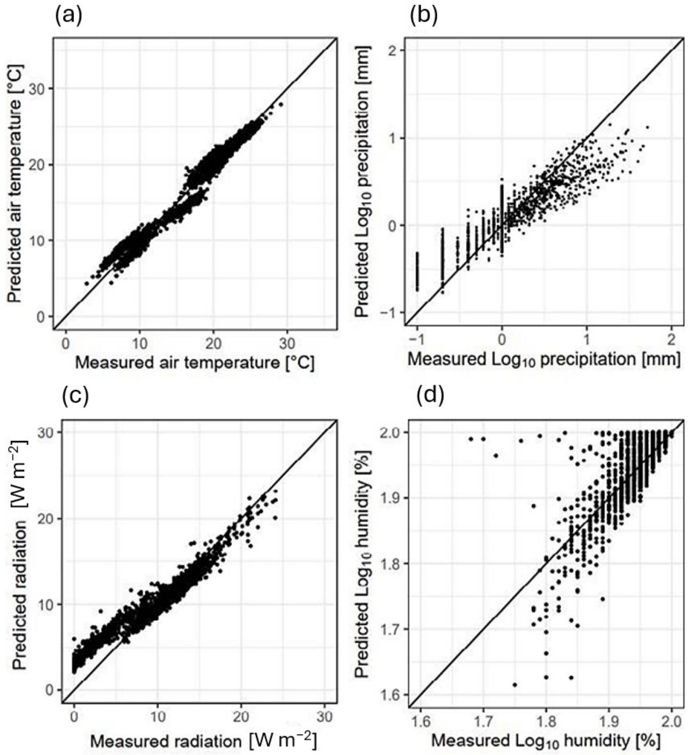

Figure 9. Results of daily microclimate models—measured vs. predicted parameters. (**a**) Mean daily air temperature in °C; (**b**) daily log10-transformed precipitation total in mm; (**c**) mean daily radiation in W m^{-2}; (**d**) mean daily log10-transformed humidity in %.

Table 2. Accuracy of the best performing models.

Parameter	Aggregation	Values	Units	RMSE	R^2
Temperature	Hourly	Absolute	°C	1.7	0.95
Precipitation	Hourly	Absolute	mm	0.09	0.53
Radiation	Hourly	Absolute	W m^{-2}	4.17	0.93
Humidity	Hourly	Log10	%	0.03	0.64
Humidity	Hourly	Absolute	%	1.07	0.64
Temperature	Daily	Absolute	°C	1.3	0.96
Precipitation	Daily	Log10	mm	0.3	0.81
Precipitation	Daily	Absolute	mm	1.9	0.81
Radiation	Daily	Absolute	W m^{-2}	1.6	0.92
Humidity	Daily	Log10	%	0.1	0.75
Humidity	Daily	Absolute	%	1.02	0.75
Temperature	Monthly	Absolute	°C	0.9	0.97
Precipitation	Monthly	Log10	mm	0.62	0.81
Precipitation	Monthly	Absolute	mm	4.9	0.81
Radiation	Monthly	Absolute	W m^{-2}	0.7	0.87
Humidity	Monthly	Log10	%	0.04	0.77
Humidity	Monthly	Absolute	%	1.09	0.77

The accuracy of our microclimate use case data compared to other similar applications is generally high, especially when its high spatio-temporal resolution is considered.

In the air temperature domain, our predictions generally outperform similar studies (Table S4) [57,61–85]. Most of their products estimate temperature at the daily temporal resolution. At this temporal resolution, our models achieved an RMSE of 1.3 °C and an R^2 of 0.96. The RMSE at the same temporal resolution was reported to be between 1.8 °C and 4.9 °C. Our hourly models reached an RMSE of 1.7 °C, surpassing the performance metrics of the most sub-daily models reported in the literature, which were reported to be between 0.31 and 11.9 °C. However, the few models outperforming our study were applied at the spatial resolution between 1 km and 4 km and not 30 m as in our case

At the monthly resolution, our models reached an RMSE of 0.97 °C, surpassing the RMSE between 1.1 °C and 2.5 °C of the previous models at a similar temporal resolution (Table S5) [86–94]. Our rainfall models reached an RMSE of 0.09 mm and R^2 of 0.53 at the hourly and an RMSE of 1.9 mm and R^2 of 0.81 at the daily temporal resolution. In comparison to other approaches , they outperform the models predicting hourly precipitation in which the RMSE stayed in the range between 1.6 and 1.8 mm. Most of the previous models predicted monthly precipitation rates only. The model performances varied strongly depending on the month (RMSE = 1.5–37.7 mm) Several studies found much worse model performance between 54.3 mm and 191.8 mm at the monthly temporal resolution. For relative air humidity, our models reached an RMSE = 1.02% and 1.09%. They strongly outperformed the models used in previous studies (Table S6) [95–101], which had an RMSE between 3.5% and 14.2%. Our radiation models performed similarly to the models in previous studies. Our estimated RMSE was between 0 and 4.14 W m^{-2}. These values can be also found in the literature, with an RMSE between 0.4 and 4.78 W m^{-2} (Table S7) [102–108]. The most common methods applied here were several types of artificial neural network models.

Figure 10 shows an example of the high-resolution microclimate raster data generated for the target variables.

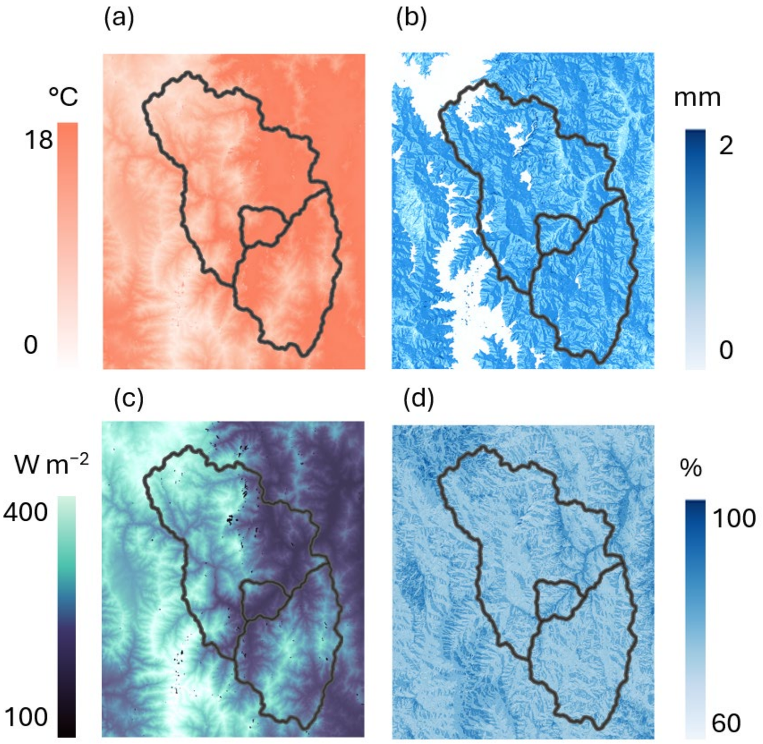

Figure 10. Spatial prediction of the parameters using the daily PLSR model for 15 April 2022 at a resolution of 30 m. (**a**) Mean average air temperature; (**b**) precipitation total; (**c**) mean average solar radiation; (**d**) mean average relative humidity. The black line indicates the MRF study domain of the research unit.

All modeled microclimate raster data are stored in the TMFdw raster engine, which results in a very high data volume. Per year, the data volume for the domain in Figure 10 with all temporal resolutions sums up to 620 GB, by far exceeding the tabular data of about 14 GB hitherto stored in total. With these data volumes, it is favorable to use a cloud platform, like the TMFdw raster engine, instead of downloading all the data and working with it locally. Given the significant data volume of 620 GB per year for raster data, compared to just 14 GB for tabular data, managing such a large dataset poses challenges. We recommend that users conduct their statistical analyses directly on the TMFdw raster engine platform rather than downloading the entire dataset locally. This approach offers several key benefits. First, it ensures that the data are not duplicated for each user, reducing storage overhead and preventing unnecessary copies. Second, by working directly on the platform, all users access the same, consistent version of the data, eliminating discrepancies caused by different data versions. Third, the platform's capabilities include caching and storage of intermediate results, which helps avoid repetitive calculations and enhances processing efficiency. Our strategy for handling these large data volumes involves centralized storage on a scalable infrastructure, minimizing the load on local devices and ensuring efficient data management. We also recognize that using the platform effectively may require additional training, particularly for ecologists who might be less familiar with this approach. To address this need, we will offer targeted training sessions, including a dedicated database workshop, to help users understand how to fully leverage the platform's features and capabilities. This approach will enable streamlined, consistent analyses while making efficient use of the available computational resources.

4. Conclusions

In this paper, we presented a comprehensive data warehouse designed for small- to medium-sized ecological research programs. The main objective of the initial development in 2002 was a meta-database for standardized data communication between the members of a joint ecological project in order to strengthen joint synergetic data evaluation [36]. The growing realization that collected research data should be open for reuse [29] led to the development of the first version of the current TMFdw data warehouse as a DOI data center [24]. The TMFdw was planned as a project database to serve as a standardized but temporary project database for ecological projects, whose data could be transferred to a final permanent repository after the end of the project period. In order to simplify the management of interdisciplinary projects, various functions (literature database, management tools for research infrastructure, news system, etc.) have also been integrated. New requirements in ecological research, such as the use of large raster datasets, especially from remote sensing and ecological modeling, as well as the increasing desire to make simple analysis functions available via machine learning, have ultimately led to the new developments presented in this paper [21,109,110].

The flexibility of the database, which is available as open source, has been demonstrated by its diverse implementation in a wide range of interdisciplinary environmental projects with heterogeneous datasets (Supplementary Materials) and shows the impact on the field of research data management in interdisciplinary environmental research programs.

The current TMFdw is a unique open-source working database for medium- and small-to-medium-scale environmental research projects and also offers a comprehensive range of services. It is easy to implement and adapt to coordinated environmental projects. A main innovation of the recent version is the inclusion of the raster data engine to include big data from remote sensing and environmental modeling, including easy-to-use visualization and analysis tools. It bridges the gap between tabular ecological data and large-scale spatio-temporal datasets. This allows researchers to enrich biological data recorded at individual research plots with respective abiotic data taken from global grid datasets for their analyses. Ecologists frequently ask for high-resolution grid data on the microclimate, which are mostly not available as grid datasets for data enrichment. The second innovation of the TMFdw is, therefore, the implementation of user-demanded use cases, such as generating and automatically updating spatio-temporally high-resolution grid data of the microclimate based on microclimate observations, data grids, such as DEMs, and machine learning methods.

The technical validation of the database is mainly based on user feedback. Reported malfunctions or useful user-requested additions are immediately implemented in the current project version and adopted in the open GitHub version. The regularly incoming logger-based data are partly checked using simple correction algorithms (for climate data, refer to [41]). Users are otherwise responsible for the quality assurance of the data they upload but must include quality information in the metadata. The generated secondary grit data of the use cases, such as the microclimate grids, are validated with independent data in the scope of the machine learning model development, and the accuracy metrics (e.g., RMSE; MAE) are reported in the quality section of the metadata.

Current limitations are mainly related to user support. There is limited capacity to support other external interdisciplinary projects during TMFdw implementation and adaptation to the project. Even if the use of the TMFdw is largely self-explanatory and is supported with help texts for the individual functions, it is recommended to at least train new users through regular user workshops. A main use note is to carefully use the data in the TMFdw. Due to the different accuracies of primary and secondary data, users must be asked to carefully study the quality of information in the metadata before using them for their own analyses. The main use note for new TMFdw operators is to think about data categories needed in their project, which have to be established before data can be uploaded. In terms of scalability, very large datasets can have high storage costs. Due to the novelty of the approach, experiences with scalability are limited so far.

In addition to the general improvement of the TMFdw and its documentation, two future directions will be pursued with priority. First, we strive to implement further use cases and enable users to run their own R or Python scripts in the TMF environment. This would significantly expand the user-defined analysis functions and thus may convert the TMFdw step by step to a comprehensive data and analysis platform. Second, we will strengthen the TMFdw by connecting it to ecological models, such as LSMs (land surface models) and Dynamic Vegetation models (DGVMs). Appropriate interfaces should make it possible to use data for model parameterization and forcing directly from the database and to save the results of the model run with their metadata directly in the DW.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/data9120143/s1, Table S1. Successful implementations of the TMFdw. Table S2. List of remote sensing and model data provided by the TMFdw raster engine for HUMBOL-TD model forcing and testing to generate future land use scenarios. Table S3. List of the use AWS data for the measuring network in the south Ecuadorian Andes. Table S4. Summary of reference articles for area-wide air temperature predictions. Table S5. Summary of reference articles for area-wide precipitation predictions. Table S6. Summary of reference articles for area-wide relative humidity predictions. Table S7. Summary of reference articles for area-wide solar radiation predictions.

Author Contributions: J.B., P.G., M.D., J.D. and C.B.: conceptualization and writing the original draft. M.D., J.D., C.B. and M.M. developed the TMFdw and its raster engine. M.D. and P.G. developed and implemented the use case. Data curation: P.S. Funding acquisition: J.B. and N.F. All authors have read and agreed to the published version of the manuscript.

Funding: The development of the TMFdw was kindly funded by the German Research Foundation (DFG) in the interdisciplinary Research Units FOR402 and FOR816 and the DFG knowledge transfer program PAK823-825. The current work, including the raster engine and the development of the use cases, was funded in the scope of the DFG research unit FOR2730 "RESPECT" in the subprojects A1 (Funding number BE1780/51-2) and Z (FA925/16-1).

Data Availability Statement: Data are available via the TMFdw (www.TropicalMountainForest.org, accessed on 3 December 2024).

Acknowledgments: The authors would like to thank the Ministerio del Ambiente, Agua y Transición Ecológica (MAATE), and the Instituto Nacional de Biodiversidad de Ecuador (INABIO) for granting research permits, and the foundation "Nature and Culture International" (NCI) for providing research facilities. They also acknowledge all of the scientists of the DFG-funded programs uploading their valuable data treasures.

Conflicts of Interest: Authors Christian Beilschmidt, Johannes Drönner and Michael Mattig were employed by the company Geo Engine GmbH. The remaining authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. https://doi.org/10.1038/sdata.2016.18.
2. Jati, P.H.P.; Lin, Y.; Nodehi, S.; Cahyono, D.B.; van Reisen, M. FAIR Versus Open Data: A Comparison of Objectives and Principles. *Data Intell.* **2022**, *4*, 867–881. https://doi.org/10.1162/dint_a_00176.
3. Hasnain, A.; Rebholz-Schuhmann, D. Assessing FAIR Data Principles Against the 5-Star Open Data Principles. In T*he Semantic Web: ESWC 2018 Satellite Events*; Gangemi, A., Gentile, A.L., Nuzzolese, A.G., Rudolph, S., Maleshkova, M., Paulheim, H., Pan, J.Z., Alam, M., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp 469–477. ISBN 978-3-319-98191-8.
4. Miyakawa, T. No raw data, no science: Another possible source of the reproducibility crisis. *Mol. Brain* **2020**, *13*, 24. https://doi.org/10.1186/s13041-020-0552-2.
5. Wittenburg, P. Open Science and Data Science. *Data Intell.* **2021**, *3*, 95–105. https://doi.org/10.1162/dint_a_00082.
6. Cao, L. A New Age of AI: Features and Futures. *IEEE Intell. Syst.* **2022**, *37*, 25–37. https://doi.org/10.1109/MIS.2022.3150944.
7. Michener, W.K.; Porter, J.; Servilla, M.; Vanderbilt, K. Long term ecological research and information management. *Ecol. Inform.* **2011**, *6*, 13–24. https://doi.org/10.1016/j.ecoinf.2010.11.005.
8. Shin, N.; Shibata, H.; Osawa, T.; Yamakita, T.; Nakamura, M.; Kenta, T. Toward more data publication of long-term ecological observations. *Ecol. Res.* **2020**, *35*, 700–707. https://doi.org/10.1111/1440-1703.12115.
9. Kaplan, N.E.; Baker, K.S.; Karasti, H. Long live the data! Embedded data management at a long-term ecological research site. *Ecosphere* **2021**, *12*, 111. https://doi.org/10.1002/ecs2.3493.

10. Lane, M.A.; Edwards, J.L. The global biodiversity information facility (GBIF). *Syst. Assoc. Spec. Vol.* **2007**, *73*, 1.
11. Diepenbroek, M.; Glöckner, F.O.; Grobe, P.; Güntsch, A.; Huber, R.; König-Ries, B.; Kostadinov, I.; Nieschulze, J.; Seeger, B.; Tolksdorf, R.; et al. Towards an integrated biodiversity and ecological research data management and archiving platform: The German federation for the curation of biological data (GFBio). *Informatik* **2014**, 1711–1721. Available online: https://dl.gi.de/server/api/core/bitstreams/8e719672-6473-4f93-83a9-d8bef7535b13/content (accessed on 1 October 2024).
12. Luther, K.; Güntsch, A.; Koenig-Ries, B.; Fichtmueller, D. NFDI4Biodiversity: A German infrastructure for biodiversity data. *Biodivers. Inf. Sci. Stand.* **2022**, *6*, e93869. https://doi.org/10.3897/biss.6.93869.
13. Ebert, B.; Engel, J.S.; Kostadinov, I.; Güntsch, A.; Glöckner, F.O. Connecting National and International Data Infrastructures in Biodiversity Research. In Proceedings of the 1st Conference on Research Data Infrastructure, Karlsruhe, Germany, 12–14 September 2023; Volume 1. https://doi.org/10.52825/cordi.v1i.346.
14. Zeuss, D.; Bald, L.; Gottwald, J.; Becker, M.; Bellafkir, H.; Bendix, J.; Bengel, P.; Beumer, L.T.; Brandl, R.; Brändle, M.; et al. Nature 4.0: A networked sensor system for integrated biodiversity monitoring. *Glob. Change Biol.* **2024**, *30*, e17056. https://doi.org/10.1111/gcb.17056.
15. Bach, K.; Schäfer, D.; Enke, N.; Seeger, B.; Gemeinholzer, B.; Bendix, J. A comparative evaluation of technical solutions for long-term data repositories in integrative biodiversity research. *Ecol. Inform.* **2012**, *11*, 16–24. https://doi.org/10.1016/j.ecoinf.2011.11.008.
16. Enke, N.; Thessen, A.; Bach, K.; Bendix, J.; Seeger, B.; Gemeinholzer, B. The user's view on biodiversity data sharing—Investigating facts of acceptance and requirements to realize a sustainable use of research data. *Ecol. Inform.* **2012**, *11*, 25–33. https://doi.org/10.1016/j.ecoinf.2012.03.004.
17. Jeppesen, J.H.; Ebeid, E.; Jacobsen, R.H.; Toftegaard, T.S. Open geospatial infrastructure for data management and analytics in interdisciplinary research. *Comput. Electron. Agric.* **2018**, *145*, 130–141. https://doi.org/10.1016/j.compag.2017.12.026.
18. Gadelha, L.M.R.; de Siracusa, P.C.; Dalcin, E.C.; da Silva, L.A.E.; Augusto, D.A.; Krempser, E.; Affe, H.M.; Costa, R.L.; Mondelli, M.L.; Meirelles, P.M.; et al. A survey of biodiversity informatics: Concepts, practices, and challenges. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2021**, *11*, 251. https://doi.org/10.1002/widm.1394.
19. Skidmore, A.K.; Coops, N.C.; Neinavaz, E.; Ali, A.; Schaepman, M.E.; Paganini, M.; Kissling, W.D.; Vihervaara, P.; Darvishzadeh, R.; Feilhauer, H.; et al. Priority list of biodiversity metrics to observe from space. *Nat. Ecol. Evol.* **2021**, *5*, 896–906. https://doi.org/10.1038/s41559-021-01451-x.
20. Wagemann, J.; Siemen, S.; Seeger, B.; Bendix, J. A user perspective on future cloud-based services for Big Earth data. *Int. J. Digit. Earth* **2021**, *14*, 1758–1774. https://doi.org/10.1080/17538947.2021.1982031.
21. Wagemann, J.; Siemen, S.; Seeger, B.; Bendix, J. Users of open Big Earth data—An analysis of the current state. *Comput. Geosci.* **2021**, *157*, 104916. https://doi.org/10.1016/j.cageo.2021.104916.
22. Wang, W.; Göpfert, T.; Stark, R. Data Management in Collaborative Interdisciplinary Research Projects—Conclusions from the Digitalization of Research in Sustainable Manufacturing. *Int. J. Geo-Inf.* **2016**, *5*, 41. https://doi.org/10.3390/ijgi5040041.
23. Garwood, D.A.; Poole, A.H. Project management as information management in interdisciplinary research: "Lots of different pieces working together". *Int. J. Inf. Manag.* **2018**, *41*, 14–22. https://doi.org/10.1016/j.ijinfomgt.2018.03.002.
24. Lotz, T.; Nieschulze, J.; Bendix, J.; Dobbermann, M.; König-Ries, B. Diverse or uniform?—Intercomparison of two major German project databases for interdisciplinary collaborative functional biodiversity research. *Ecol. Inform.* **2012**, *8*, 10–19. https://doi.org/10.1016/j.ecoinf.2011.11.004.
25. Chamanara, J.; Gaikwad, J.; Gerlach, R.; Algergawy, A.; Ostrowski, A.; König-Ries, B. BEXIS2: A FAIR-aligned data management system for biodiversity, ecology and environmental data. *Biodivers. Data J.* **2021**, *9*, e72901. https://doi.org/10.3897/BDJ.9.e72901.
26. Beck, E.; Müller-Hohenstein, K. Analysis of undisturbed and disturbed tropical mountain forest ecosystems in Southern Ecuador. *Erde* **2001**, *132*, 1–8.
27. Beck, E. *Gradients in a Tropical Mountain Ecosystem of Ecuador*; Springer Berlin/Heidelberg, Germany, 2008; ISBN 978-3-540-73525-0.
28. Bendix, J.; Beck, E.; Bräuning, A.; Makeschin, F.; Mosandl, R.; Scheu, S.; Wilcke, W. *Ecosystem Services, Biodiversity and Environmental Change in a Tropical Mountain Ecosystem of South Ecuador*; Beck, E., Bräuning, A., Makeschin, F., Mosandl, R., Scheu, S., Wilcke, W., Bendix, J., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; ISBN 978-3-642-38136-2.
29. Bendix, J.; Nieschulze, J.; Michener, W.K. Data platforms in integrative biodiversity research. *Ecol. Inform.* **2012**, *11*, 1–4. https://doi.org/10.1016/j.ecoinf.2012.04.001.
30. Beck, E.; Knoke, T.; Farwig, N.; Breuer, L.; Siddons, D.; Bendix, J. *Landscape Restoration, Sustainable Land Use and Cross-Scale Monitoring of Biodiversity and Ecosystem Functions. A Science-Directed Approach for South Ecuador*; Universität Bayreuth: Bayreuth, Germany, 2017. https://doi.org/10.5678/lcrs/pak823-825.cit.1696.
31. Farwig, N.; Bendix, J.; Beck, E. Introduction to the Special Issue "Functional monitoring in megadiverse tropical ecosystems". *Ecol. Indic.* **2017**, *83*, 524–526. https://doi.org/10.1016/j.ecolind.2017.02.027.
32. Bendix, J.; Rollenbeck, R.; Göttlicher, D.; Cermak, J. Cloud occurrence and cloud properties in Ecuador. *Clim. Res.* **2006**, *30*, 133–147. https://doi.org/10.3354/cr030133.
33. Bendix, J.; Rollenbeck, R.; Palacios, W.E. Cloud detection in the Tropics--a suitable tool for climate-ecological studies in the high mountains of Ecuador. *Int. J. Remote Sens.* **2004**, *25*, 4521–4540. https://doi.org/10.1080/01431160410001709967.
34. Environmental Changes in Biodiversity Hotspot Ecosystems of South Ecuador: RESPonse and Feedback effECTs (FOR2730). Available online: https://vhrz669.hrz.uni-marburg.de/tmf_respect/ (accessed on 26 November 2024).

35. Bendix, J.; Aguire, N.; Beck, E.; Bräuning, A.; Brandl, R.; Breuer, L.; Böhning-Gaese, K.; de Paula, M.D.; Hickler, T.; Homeier, J.; et al. A research framework for projecting ecosystem change in highly diverse tropical mountain ecosystems. *Oecologia* **2021**, *195*, 589–600. https://doi.org/10.1007/s00442-021-04852-8.
36. Göttlicher, D.; Bendix, J. Eine modulare Multi-User Datenbank für eine ökologische Forschergruppe mit heterogenem Datenbestand (A modular multi-user database for an ecological research group with a heterogeneous database). *Z. Agrar.* **2004**, *4*, 95–103. Available online: https://www.researchgate.net/publication/284297182_Eine_modulare_Multi-User_Datenbank_fur_eine_okologische_Forschergruppe_mit_heterogenem_Datenbestand (accessed on 26 November 2024).
37. GitLab. LCRS Marburg. Available online: https://gitlab.com/lcrsmarburg (accessed on 26 November 2024).
38. Pampel, H.; Vierkant, P.; Scholze, F.; Bertelmann, R.; Kindling, M.; Klump, J.; Goebelbecker, H.-J.; Gundlach, J.; Schirmbacher, P.; Dierolf, U. Making research data repositories visible: The re3data.org Registry. *PLoS ONE* **2013**, *8*, e78080. https://doi.org/10.1371/journal.pone.0078080.
39. RESPECT Repository. Available online: http://doi.org/10.17616/R3VF82 (accessed on 26 November 2024).
40. RESPECT Data Agreement. Available online: https://vhrz669.hrz.uni-marburg.de/tmf_respect/UserFiles/File/respect/generalinformations/RESPECT_data_use_agreement_approved.pdf (accessed on 26 November 2024).
41. Fegraus, E.H.; Andelman, S.; Jones, M.B.; Schildhauer, M. Maximizing the value of ecological data with structured metadata: An introduction to ecological metadata language (EML) and principles for metadata creation. *Bull. Ecol. Soc. Am.* **2005**, *86*, 158–168.
42. Ecological Metadata Language Version 2.2.0; KNB Data Repository. 2019. Available online: https://eml.ecoinformatics.org/ (accessed on 26 November 2024).
43. Wieczorek, J.; Bloom, D.; Guralnick, R.; Blum, S.; Döring, M.; Giovanni, R.; Robertson, T.; Vieglais, D. Darwin Core: An evolving community-developed biodiversity data standard. *PLoS ONE* **2012**, *7*, e29715. https://doi.org/10.1371/journal.pone.0029715.
44. Rollenbeck, R.; Trachte, K.; Bendix, J. A New Class of Quality Controls for Micrometeorological Data in Complex Tropical Environments. *J. Atmos. Ocean. Technol.* **2016**, *33*, 169–183. https://doi.org/10.1175/JTECH-D-15-0062.1.
45. Dantas de Paula, M.; Forrest, M.; Langan, L.; Bendix, J.; Homeier, J.; Velescu, A.; Wilcke, W.; Hickler, T. Nutrient cycling drives plant community trait assembly and ecosystem functioning in a tropical mountain biodiversity hotspot. *New Phytol.* **2021**, *232*, 551–566. https://doi.org/10.1111/nph.17600.
46. Beilschmidt, C.; Drönner, J.; Mattig, M.; Schweitzer, P.; Seeger, B. Geo Engine: Workflow-Backed Geo Data Portals. 2023. Available online: https://dl.gi.de/server/api/core/bitstreams/c549a4d3-b898-436d-b664-a9cdc6fec6ba/content (accessed on 26 November 2024).
47. Sakimura, N.; Bradley, J.; Jones, M.; de Medeiros, B.; Mortimore, C. OpenID Connect Core 1.0. 2023. Available online: https://openid.net/specs/openid-connect-core-1_0.html (accessed on 26 November 2024).
48. Thorgersen, S.; Silva, P.I. *Keycloak—Identity and Access Management for Modern Applications: Harness the Power of Keycloak, OpenID Connect, and OAuth 2.0 Protocols to Secure Applications*; Packt Publishing: Birmingham, UK, 2021; ISBN 978-1-80056-249-3.
49. RESPECT Reasearch Group. Available online: https://vhrz669.hrz.uni-marburg.de/tmf_respect/content_projects.do?phase=5&subpage=staff (accessed on 26 November 2024).
50. DEM Ecuador. Available online: http://www.tropicalmountainforest.org/data_pre.do?citid=1400 (accessed on 26 November 2024).
51. Fleming, M.D.; Hoffer, R.M. Machine Processing of Landsat MSS Data and DMA Topographic Data for Forest Cover Type Mapping. LARS Technical Report 062879. 1979. Available online: https://docs.lib.purdue.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1079&context=larstech (accessed on 26 November 2024).
52. Ritter, P. A vector-based slope and aspect generation algorithm. *Photogramm. Eng. Remote Sens.* **1987**, *53*, 1109–1111.
53. Wilson, M.F.J.; O'Connell, B.; Brown, C.; Guinan, J.C.; Grehan, A.J. Multiscale Terrain Analysis of Multibeam Bathymetry Data for Habitat Mapping on the Continental Slope. *Mar. Geod.* **2007**, *30*, 3–35. https://doi.org/10.1080/01490410701295962.
54. Obermeier, W.A.; Lehnert, L.W.; Pohl, M.J.; Makowski Gianonni, S.; Silva, B.; Seibert, R.; Laser, H.; Moser, G.; Müller, C.; Luterbacher, J.; et al. Grassland ecosystem services in a changing environment: The potential of hyperspectral monitoring. *Remote Sens. Environ.* **2019**, *232*, 111273. https://doi.org/10.1016/j.rse.2019.111273.
55. Ringle, C.M.; Sarstedt, M.; Sinkovics, N.; Sinkovics, R.R. A perspective on using partial least squares structural equation modelling in data articles. *Data Brief* **2023**, *48*, 109074. https://doi.org/10.1016/j.dib.2023.109074.
56. Göttlicher, D.; Obregón, A.; Homeier, J.; Rollenbeck, R.; Nauss, T.; Bendix, J. Land-cover classification in the Andes of southern Ecuador using Landsat ETM+ data as a basis for SVAT modelling. *Int. J. Remote Sens.* **2009**, *30*, 1867–1886. https://doi.org/10.1080/01431160802541531.
57. Fries, A.; Rollenbeck, R.; Göttlicher, D.; Nauß, T.; Homeier, J.; Peters, T.; Bendix, J. Thermal structure of a megadiverse Andean mountain ecosystem in southern Ecuador and its regionalization. *Erdkunde* **2009**, *63*, 321–335. https://doi.org/10.3112/erdkunde.2009.04.03.
58. Fries, A.; Rollenbeck, R.; Nauß, T.; Peters, T.; Bendix, J. Near surface air humidity in a megadiverse Andean mountain ecosystem of southern Ecuador and its regionalization. *Agric. For. Meteorol.* **2012**, *152*, 17–30. https://doi.org/10.1016/j.agrformet.2011.08.004.
59. Bendix, A.; Bendix, J. Heavy rainfall episodes in Ecuador during El Niño events and associated regional atmospheric circulation and SST patterns. *Adv. Geosci.* **2006**, *6*, 43–49. https://doi.org/10.5194/adgeo-6-43-2006.

60. Onyutha, C. From R-Squared to Coefficient of Model Accuracy for Assessing "Goodness-of-Fits". 2020. Available online: https://doi.org/10.5194/gmd-2020-51 (accessed on 26 November 2024).
61. Cristóbal, J.; Ninyerola, M.; Pons, X. Modeling air temperature through a combination of remote sensing and GIS data. *J. Geophys. Res.* **2008**, *113*, 55. https://doi.org/10.1029/2007JD009318.
62. Chen, F.; Liu, Y.; Liu, Q.; Qin, F. A statistical method based on remote sensing for the estimation of air temperature in China. *Int. J. Climatol.* **2015**, *35*, 2131–2143. https://doi.org/10.1002/joc.4113.
63. Golkar, F.; Sabziparvar, A.A.; Khanbilvardi, R.; Nazemosadat, M.J.; Zand-Parsa, S.; Rezaei, Y. Estimation of instantaneous air temperature using remote sensing data. *Int. J. Remote Sens.* **2018**, *39*, 258–275. https://doi.org/10.1080/01431161.2017.1382743.
64. Liu, S.; Su, H.; Tian, J.; Zhang, R.; Wang, W.; Wu, Y. Evaluating Four Remote Sensing Methods for Estimating Surface Air Temperature on a Regional Scale. *J. Appl. Meteorol. Climatol.* **2017**, *56*, 803–814. https://doi.org/10.1175/JAMC-D-16-0188.1.
65. Jang, K.; Kang, S.; Kimball, J.; Hong, S. Retrievals of All-Weather Daily Air Temperature Using MODIS and AMSR-E Data. *Remote Sens.* **2014**, *6*, 8387–8404. https://doi.org/10.3390/rs6098387.
66. Zhu, W.; Lű, A.; Jia, S. Estimation of daily maximum and minimum air temperature using MODIS land surface temperature products. *Remote Sens. Environ.* **2013**, *130*, 62–73. https://doi.org/10.1016/j.rse.2012.10.034.
67. Pape, R.; Löffler, J. Modelling spatio-temporal near-surface temperature variation in high mountain landscapes. *Ecol. Model.* **2004**, *178*, 483–501. https://doi.org/10.1016/j.ecolmodel.2004.02.019.
68. Hou, P.; Chen, Y.; Qiao, W.; Cao, G.; Jiang, W.; Li, J. Near-surface air temperature retrieval from satellite images and influence by wetlands in urban region. *Theor. Appl. Climatol.* **2013**, *111*, 109–118. https://doi.org/10.1007/s00704-012-0629-7.
69. Shen, H.; Jiang, Y.; Li, T.; Cheng, Q.; Zeng, C.; Zhang, L. Deep learning-based air temperature mapping by fusing remote sensing, station, simulation and socioeconomic data. *Remote Sens. Environ.* **2020**, *240*, 111692. https://doi.org/10.1016/j.rse.2020.111692.
70. Hooker, J.; Duveiller, G.; Cescatti, A. A global dataset of air temperature derived from satellite remote sensing and weather stations. *Sci. Data* **2018**, *5*, 180246. https://doi.org/10.1038/sdata.2018.246.
71. Şahin, M. Modelling of air temperature using remote sensing and artificial neural network in Turkey. *Adv. Space Res.* **2012**, *50*, 973–985. https://doi.org/10.1016/j.asr.2012.06.021.
72. Zhang, Z.; Du, Q. A Bayesian Kriging Regression Method to Estimate Air Temperature Using Remote Sensing Data. *Remote Sens.* **2019**, *11*, 767. https://doi.org/10.3390/rs11070767.
73. Hadria, R.; Benabdelouahab, T.; Mahyou, H.; Balaghi, R.; Bydekerke, L.; El Hairech, T.; Ceccato, P. Relationships between the three components of air temperature and remotely sensed land surface temperature of agricultural areas in Morocco. *Int. J. Remote Sens.* **2018**, *39*, 356–373. https://doi.org/10.1080/01431161.2017.1385108.
74. Li, L.; Zha, Y. Estimating monthly average temperature by remote sensing in China. *Adv. Space Res.* **2019**, *63*, 2345–2357. https://doi.org/10.1016/j.asr.2018.12.039.
75. Liu, S.; Su, H.; Zhang, R.; Tian, J.; Wang, W. Estimating the Surface Air Temperature by Remote Sensing in Northwest China Using an Improved Advection-Energy Balance for Air Temperature Model. *Adv. Meteorol.* **2016**, *2016*, 4294219. https://doi.org/10.1155/2016/4294219.
76. Xu, Y.; Knudby, A.; Ho, H.C. Estimating daily maximum air temperature from MODIS in British Columbia, Canada. *Int. J. Remote Sens.* **2014**, *35*, 8108–8121. https://doi.org/10.1080/01431161.2014.978957.
77. Kloog, I.; Nordio, F.; Coull, B.A.; Schwartz, J. Predicting spatiotemporal mean air temperature using MODIS satellite surface temperature measurements across the Northeastern USA. *Remote Sens. Environ.* **2014**, *150*, 132–139. https://doi.org/10.1016/j.rse.2014.04.024.
78. Kim, D.-Y.; Han, K.-S. Remotely sensed retrieval of midday air temperature considering atmospheric and surface moisture conditions. *Int. J. Remote Sens.* **2013**, *34*, 247–263. https://doi.org/10.1080/01431161.2012.712235.
79. Gholamnia, M.; Alavipanah, S.K.; Darvishi Boloorani, A.; Hamzeh, S.; Kiavarz, M. Diurnal Air Temperature Modeling Based on the Land Surface Temperature. *Remote Sens.* **2017**, *9*, 915. https://doi.org/10.3390/rs9090915.
80. Meyer, H.; Katurji, M.; Appelhans, T.; Müller, M.; Nauss, T.; Roudier, P.; Zawar-Reza, P. Mapping Daily Air Temperature for Antarctica Based on MODIS LST. *Remote Sens.* **2016**, *8*, 732. https://doi.org/10.3390/rs8090732.
81. Benali, A.; Carvalho, A.C.; Nunes, J.P.; Carvalhais, N.; Santos, A. Estimating air surface temperature in Portugal using MODIS LST data. *Remote Sens. Environ.* **2012**, *124*, 108–121. https://doi.org/10.1016/j.rse.2012.04.024.
82. Samanta, S.; Pal, D.K.; Lohar, D.; Pal, B. Interpolation of climate variables and temperature modeling. *Theor. Appl. Climatol.* **2012**, *107*, 35–45. https://doi.org/10.1007/s00704-011-0455-3.
83. Wang, M.; He, G.; Zhang, Z.; Wang, G.; Zhang, Z.; Cao, X.; Wu, Z.; Liu, X. Comparison of Spatial Interpolation and Regression Analysis Models for an Estimation of Monthly Near Surface Air Temperature in China. *Remote Sens.* **2017**, *9*, 1278. https://doi.org/10.3390/rs9121278.
84. Nikoloudakis, N.; Stagakis, S.; Mitraka, Z.; Kamarianakis, Y.; Chrysoulakis, N. Spatial interpolation of urban air temperatures using satellite-derived predictors. *Theor. Appl. Climatol.* **2020**, *141*, 657–672. https://doi.org/10.1007/s00704-020-03230-3.
85. Ruiz-Álvarez, M.; Alonso-Sarria, F.; Gomariz-Castillo, F. Interpolation of Instantaneous Air Temperature Using Geographical and MODIS Derived Variables with Machine Learning Techniques. *Int. J. Geo-Inf.* **2019**, *8*, 382. https://doi.org/10.3390/ijgi8090382.
86. Jing, W.; Yang, Y.; Yue, X.; Zhao, X. A Comparison of Different Regression Algorithms for Downscaling Monthly Satellite-Based Precipitation over North China. *Remote Sens.* **2016**, *8*, 835. https://doi.org/10.3390/rs8100835.

87. Kang, L.; Di, L.; Deng, M.; Shao, Y.; Yu, G.; Shrestha, R. Use of Geographically Weighted Regression Model for Exploring Spatial Patterns and Local Factors Behind NDVI-Precipitation Correlation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 4530–4538. https://doi.org/10.1109/JSTARS.2014.2361128.
88. Shen, J.; Liu, P.; Xia, J.; Zhao, Y.; Dong, Y. Merging Multisatellite and Gauge Precipitation Based on Geographically Weighted Regression and Long Short-Term Memory Network. *Remote Sens.* **2022**, *14*, 3939. https://doi.org/10.3390/rs14163939.
89. Hu, D.; Shu, H.; Hu, H.; Xu, J. Spatiotemporal regression Kriging to predict precipitation using time-series MODIS data. *Cluster Comput.* **2017**, *20*, 347–357. https://doi.org/10.1007/s10586-016-0708-0.
90. Bostan, P.A.; Heuvelink, G.B.M.; Akyurek, S.Z. Comparison of regression and kriging techniques for mapping the average annual precipitation of Turkey. *Int. J. Appl. Earth Obs. Geoinf.* **2012**, *19*, 115–126. https://doi.org/10.1016/j.jag.2012.04.010.
91. Moraux, A.; Dewitte, S.; Cornelis, B.; Munteanu, A. A Deep Learning Multimodal Method for Precipitation Estimation. *Remote Sens.* **2021**, *13*, 3278. https://doi.org/10.3390/rs13163278.
92. Bajat, B.; Pejović, M.; Luković, J.; Manojlović, P.; Ducić, V.; Mustafić, S. Mapping average annual precipitation in Serbia (1961–1990) by using regression kriging. *Theor. Appl. Climatol.* **2013**, *112*, 1–13. https://doi.org/10.1007/s00704-012-0702-2.
93. Lu, X.; Li, J.; Liu, Y.; Li, Y.; Huo, H. Quantitative Precipitation Estimation in the Tianshan Mountains Based on Machine Learning. *Remote Sens.* **2023**, *15*, 3962. https://doi.org/10.3390/rs15163962.
94. Varouchakis, E.A.; Kamińska-Chuchmała, A.; Kowalik, G.; Spanoudaki, K.; Graña, M. Combining Geostatistics and Remote Sensing Data to Improve Spatiotemporal Analysis of Precipitation. *Sensors* **2021**, *21*, 3132. https://doi.org/10.3390/s21093132.
95. Hurter, F.; Maier, O. Tropospheric profiles of wet refractivity and humidity from the combination of remote sensing data sets and measurements on the ground. *Atmos. Meas. Tech.* **2013**, *6*, 3083–3098. https://doi.org/10.5194/amt-6-3083-2013.
96. Yan, D.; Liu, T.; Dong, W.; Liao, X.; Luo, S.; Wu, K.; Zhu, X.; Zheng, Z.; Wen, X. Integrating remote sensing data with WRF model for improved 2-m temperature and humidity simulations in China. *Dyn. Atmos. Oceans* **2020**, *89*, 101127. https://doi.org/10.1016/j.dynatmoce.2019.101127.
97. Cai, X.; Bao, Y.; Petropoulos, G.P.; Lu, F.; Lu, Q.; Zhu, L.; Wu, Y. Temperature and Humidity Profile Retrieval from FY4-GIIRS Hyperspectral Data Using Artificial Neural Networks. *Remote Sens.* **2020**, *12*, 1872. https://doi.org/10.3390/rs12111872.
98. Che, Y.; Ma, S.; Xing, F.; Li, S.; Dai, Y. An improvement of the retrieval of temperature and relative humidity profiles from a combination of active and passive remote sensing. *Meteorol. Atmos. Phys.* **2019**, *131*, 681–695. https://doi.org/10.1007/s00703-018-0588-3.
99. Jiang, J.H.; Yue, Q.; Su, H.; Kangaslahti, P.; Lebsock, M.; Reising, S.; Schoeberl, M.; Wu, L.; Herman, R.L. Simulation of Remote Sensing of Clouds and Humidity From Space Using a Combined Platform of Radar and Multifrequency Microwave Radiometers. *Earth Space Sci.* **2019**, *6*, 1234–1243. https://doi.org/10.1029/2019EA000580.
100. Jackson, D.L.; Wick, G.A.; Bates, J.J. Near-surface retrieval of air temperature and specific humidity using multisensor microwave satellite observations. *J. Geophys. Res.* **2006**, *111*, 755. https://doi.org/10.1029/2005JD006431.
101. Polyakov, A.; Virolainen, Y.; Nerobelov, G.; Timofeyev, Y.; Solomatnikova, A. Total ozone measurements using IKFS-2 spectrometer aboard Meteor-M N2 satellite in 2019–2020. *Int. J. Remote Sens.* **2021**, *42*, 8709–8733. https://doi.org/10.1080/01431161.2021.1985741.
102. Şenkal, O. Modeling of solar radiation using remote sensing and artificial neural network in Turkey. *Energy* **2010**, *35*, 4795–4801. https://doi.org/10.1016/j.energy.2010.09.009.
103. Robles-Zazueta, C.A.; Molero, G.; Pinto, F.; Foulkes, M.J.; Reynolds, M.P.; Murchie, E.H. Field-based remote sensing models predict radiation use efficiency in wheat. *J. Exp. Bot.* **2021**, *72*, 3756–3773. https://doi.org/10.1093/jxb/erab115.
104. Ghimire, S.; Deo, R.C.; Downs, N.J.; Raj, N. Self-adaptive differential evolutionary extreme learning machines for long-term solar radiation prediction with remotely-sensed MODIS satellite and Reanalysis atmospheric products in solar-rich cities. *Remote Sens. Environ.* **2018**, *212*, 176–198. https://doi.org/10.1016/j.rse.2018.05.003.
105. Ghimire, S.; Deo, R.C.; Downs, N.J.; Raj, N. Global solar radiation prediction by ANN integrated with European Centre for medium range weather forecast fields in solar rich cities of Queensland Australia. *J. Clean. Prod.* **2019**, *216*, 288–310. https://doi.org/10.1016/j.jclepro.2019.01.158.
106. Yan, G.; Wang, T.; Jiao, Z.; Mu, X.; Zhao, J.; Chen, L. Topographic radiation modeling and spatial scaling of clear-sky land surface longwave radiation over rugged terrain. *Remote Sens. Environ.* **2016**, *172*, 15–27. https://doi.org/10.1016/j.rse.2015.10.026.
107. Samani, Z.; Bawazir, A.S.; Bleiweiss, M.; Skaggs, R.; Tran, V.D. Estimating Daily Net Radiation over Vegetation Canopy through Remote Sensing and Climatic Data. *J. Irrig. Drain Eng.* **2007**, *133*, 291–297. https://doi.org/10.1061/(ASCE)0733-9437(2007)133:4(291).
108. Whitlock, C.H.; Charlock, T.P.; Staylor, W.F.; Pinker, R.T.; Laszlo, I.; Ohmura, A.; Gilgen, H.; Konzelman, T.; DiPasquale, R.C.; Moats, C.D.; et al. First Global WCRP Shortwave Surface Radiation Budget Dataset. *Bull. Amer. Meteor. Soc.* **1995**, *76*, 905–922. https://doi.org/10.1175/1520-0477(1995)076<0905:FGWSSR>2.0.CO;2.
109. Pereira, H.M.; Ferrier, S.; Walters, M.; Geller, G.N.; Jongman, R.H.G.; Scholes, R.J.; Bruford, M.W.; Brummitt, N.; Butchart, S.H.M.; Cardoso, A.C.; et al. Ecology. Essential biodiversity variables. *Science* **2013**, *339*, 277–278. https://doi.org/10.1126/science.1229931.
110. Coops, N.C.; Skidmore, A. SectorInsights.com—Essential Biodiversity Variables (EBVs) and Earth Observation—An Invitation to Participate. *Photogramm. Eng. Remote Sens.* **2021**, *87*, 792–793. https://doi.org/10.14358/PERS.87.11.792.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

SUPPLEMENTARY MATERIALS

Table S1. Successful implementations of the TMFdw. FOR: DFG Research Unit. PAK: DFG bundle program, DFG: German Research Foundation, SP: subproject, LOEWE: Hessian state offensive for the development of scientific and economic excellence, D: Germany, EC: Ecuador, BMBF: Federal Ministry of Education and Research. Current development is mainly funded by the DFG RU2730 RESPECT (bold). For further information, refer to (https://uni-marburg.de/EjXEhf). *over three 2-years funding phases. **Data from long-term research in the biodiversity hotspot of the SE-Andes in Ecuador included in the TMFdw.

Program	Topic	Duration	No SP	Funding Agency	URL
TMFdw					
**PK EC	Biodiversity, inventory	1997-2001	11	DFG	-
**FOR402	Biodiversity, functioning	2001-2005	28*	DFG	http://bergregenwald.de/pages/02About.html
**FOR816	Biodiversity, ecosystem services	2006-2013	25 + 4 BMBF + 5 EC	DFG	https://vhrz669.hrz.uni-marburg.de/tmf_respect/content_projects.do?phase=2&subpage=intro
**PAK823-825	Knowledge Transfer	2014-2018	17D + 4EC	DFG	https://vhrz669.hrz.uni-marburg.de/tmf_respect/content_projects.do?phase=3&subpage=intro
**FOR2730	Biodiversity, area-wide modelling, RS, AI	Since 2018	11	DFG	https://vhrz669.hrz.uni-marburg.de/tmf_respect/home.do
Further	**Implementations**				**Further implementations**
FACE2FACE	Climate change (FACE), carbon cycle	2014-2017	13	LOEWE	www.face2face.center
CorsicArchive	Dendro-climatology	2017-2020	4	DFG	www.CorsicArchive.de

FOR2358	Landscape history	2017-2023	8	DFG	https://www.uni-marburg.de/en/fb19/dfg2358
FOR2337	Nitrogen cycle	2016-2022	9	DFG	-
FOR5288	Hydrology - Stormflow	Since 2020	8	DFG	http://ssf-hydrology.org
DARWIN	Climatology, AI	Since 2022-	2	DFG	www.darwin-rain.org
HABITAT	Health-climate change, AI	Since 2024	12	LOEWE	https://vhrz669.hrz.uni-marburg.de/habitat/)

Table S2. List of remote sensing and model data provided by the TMFdw raster engine for HUMBOL-TD model forcing and testing, and to generate future land use scenarios. LST: Land Surface Temperature, ET: Evapotranspiration

Dataset	Description	Resolution	Source	Integration
Sentinel-2	Level 2C from Copernicus	10-60 m	Element 84 STAC	A
ALOS PALSAR	Digital Elevation Model	30 m	NASA Earthdata	SA
ASTER	Digital Elevation Map	30 m	NASA Earthdata	SA
CMIP6	Future climate scenarios		ECMWF Climate Data Store	SA
ERA5	reanalysis data with multiple climate variables: - air temperature - eastward wind - northward wind - specific humidity - relative humidity	0.25°	ECMWF Climate Data Store	SA
ERA5 Land	reanalysis data with multiple climate variables: - 2m temperature - 10m u component of wind - 10m v component of wind - 2m dewpoint temperature - surface net solar radiation - surface pressure - surface solar radiation downwards - surface thermal radiation downwards - total precipitation	0.1°	ECMWF Climate Data Store	SA
ECOSTRESS	Derived products for canopy and soil: - ETcanopy - ETdaily - ETinst - ETinstUncertainty - ETinterception - ETsoil	70 m	NASA Earthdata	SA
Sentinel-3 LAI	Leaf Area Index	333 m	Copernicus WCS service by Vito	SA
MODIS	Monthly EVI & NDVI	0.05°	USGS	SA

TPI	Landform data, aspect & slope	30 m		SA
SRTM	Digital elevation model	30 m	NASA Earthdata	M

Table S3. List of use AWS data for the measuring network in the south Ecuadorian Andes (Cit-ID is http://www.lcrs.de/data_pre.do?citid=*xxxx*).

Station	**Latitude**	**Longitude**	**Elevation**	**Land cover**	**Cit-ID**
			[m a.s.l.]		***xxxx***
Bombuscaro	-4.11444	-78.9650	1234	Pasture	1712
Bombuscaro Bosque	-4.11526	-78.9673	1173	Forest	1888
Cajunama	-4.11414	-79.1751	2749	Subparamo	2003
Cajunama Pastos	-4.08976	-79.1846	2749	Pasture	1887
Cajunama Bosque	-4.11445	-79.1750	2732	Forest	1886
ECSF	-3.97254	-79.0763	2033	Forest	1713
ECSF Pastos	-3.96684	-79.0753	1957	Pasture	1718
ECSF Pinus	-3.96842	-79.0793	2004	Forest	1747
ECSF Bosque	-3.97351	-79.0761	1826	Forest	2021
El Tiro	-3.97917	-79.1439	2825	Subparamo	1714
Zamora Pastos	-3.62223	-78.6825	1338	Pasture	1847

Table S4. Summary of reference articles for area-wide air temperature predictions

Reference	Method	RMSE [°C]	R^2	Spatial resolution	Temporal resolution
[1]	Random forest	1.7	0.60	120 m	Day
[1]	Random forest	0.8	0.84	120 m	Month
[2]	Regression	1.2 – 1.4	0.95 – 0.99	5 km	Month
[3]	Extrapolation	4.9	0.71	250 m	Diurnal
[4]	Regression	0.6 – 0.8	0.70 – 0.74	90 m	2 days
[5]	Linear regression	1.8	0.93	5 km	Day
[6]	TVX algorithm	3.4	-	1 km	Day
[7]	Energy balance	2.1	0.90	10 km	Hour
[8]	Bowen equilibrium	2.2	-	1 km	Day
[9]	Deep learning	1.9 – 3.6	0.94 – 0.98	1 km	Month
[10]	Regression	1.5 – 2.1	-	500 km	Month
[11]	Neural network	1.2	0.99	1 km	Month
[12]	Regression	1.2	0.90	5 km	Month
[13]	Linear regression	2.4 – 3.9	0.59 – 0.79	1 km	Month
[14]	Regression	1.1 – 1.4	-	1 km	Month
[15]	Energy balance	0.3	0.77	90 m	Month

[16]	Random forest	-	0.74	1 km	Month
[6]	Linear regression	2.9	0.88	1 km	Month
[17]	Regression	2.1 – 2.3	0.94	1 km	Day
[18]	Regression	3.0	-	1 km	3 months
[19]	Regression	2.1 – 2.4	-	4 km	Day
[20]	Random forest	10.5 – 11.9	0.56 – 0.71	1 km	Hour
[21]	Boostrapping	1.3	0.94	1 km	Day
[22]	Interpolation	-	-	50 km	Day
[23]	Interpolation	1.5 – 2.3	-	1 km	Month
[24]	Interpolation	0.4	-	10 km	Year
[25]	interpolation	3.0	0.88	5 km	Month
[26]	Interpolation	-	-	0.1 km	Month

Table S5. Summary of reference articles for area-wide precipitation predictions

Reference	Method	RMSE [mm]	R^2	Spatial resolution	Temporal resolution
[27]	Regression	6.0 - 38	0.34 – 0.57	1 km	Month
	Random forest	0.6 – 6.2	0.98	1 km	Month
	CART	1.0 – 14.1	0.91 – 0.98	1 km	Month
	k-NN	2.2 – 19.4	0.79 – 0.94	1 km	Month
	SVM	2.9 – 37.7	0.71 – 0.88	1 km	Month
[28]	Ordinary least square	-	0.56 – 0.77	250 m	Day
	Spatial log model	-	0.62 – 0.89	250 m	Day
	Regression	-	0.74 – 0.93	250 m	Day
[29]	Regression	4.55	0.86	10 km	Month
[30]	Log-linear kriging	12.2	0.42	1 km	Month
	Log-space kriging	13.8	0.37	1 km	Month
[31]	Universal kriging	178	0.61	5 km	Year
	Ordinal kriging	201 - 211	0.45 – 0.5	5 km	Year
	Regression kriging	186	0.57	5 km	Year
	Linear regression	222	0.39	5 km	Year
[32]	Deep learning	1.6 – 1.8	-	2 km	10 min
[33]	Regression kriging	-	0.84	1 km	Year

[29]	Random forest	1.5 – 2.1	0.92	5 km	10 min
[34]	Stepwise regression	28.11	-	10 km	Month
	Weighted regression	20.94	-	10 km	Month
	Random forest	19.81	-	10 km	Month
[35]	Space-time kriging	59.30	-	5 km	Year

Table S6. Summary of reference articles for area-wide relative humidity predictions

Reference	**Method**	**RMSE [%]**	**R^2**	**Spatial resolution**	**Temporal resolution**
[36]	Least-squares collocation	7.3	-	3 km	Day
[37]	WRF model	13.2 – 19.1	0.73 – 0.92	25 km	6 hours
[38]	Neural network	5.1	0.98	16 km	6 hours
[39]	Neural network	16.0	-	10 km	Day
[40]	Bayesian Retrieval	20 - 30	-	16 km	Day
[41]	Multivariate regression	3.5 – 5.8	0.80 – 0.96	30 km	Day
[42]	Neural network	10-15	-	30 km	Hour
	Multivariate regression	2-4	-	30 km	Hour
	IPMA	2-4	-	30 km	Hour

Table S7. Summary of reference articles for area-wide solar radiation predictions

Reference	Method	RMSE	R^2	Spatial resolution	Temporal resolution
[43]	Neural network	0.16 – 0.32	0.93 – 0.95	1.1 km	Day
[44]	Partial-least square regression	4.78	0.70 – 0.93	0.5 m	Day
[45]	Extreme learning machine	0.405 – 0.645	0.96 – 0.99	25 km	Month
[46]	Neural network	1.613 - 2.275	0.86 – 0.92	80 km	Day
	Support vector machine	1.994 – 2.589	0.84 – 0.90	80 km	Day
	Gaussian process learning	2.065 – 2.723	0.82 – 0.88	80 km	Day
	Genetic programming	2.142 – 2.856	0.79 – 0.88	80 km	Day
[47]	Neural network	23	-	5 km	Day
[48]	Regression	1.32 – 2.66	-	60 km	8 days
	Regression	1.25 – 3.20	-	60 km	Day
[49]	Pinker algorithm	1.06 – 5.34	-	280 km	Day
	Pinker algorithm	5.34	-	280 km	Month

References

1. Cristóbal, J.; Ninyerola, M.; Pons, X. Modeling air temperature through a combination of remote sensing and GIS data. *J. Geophys. Res.* **2008**, *113*, 55, doi:10.1029/2007JD009318.

2. Chen, F.; Liu, Y.; Liu, Q.; Qin, F. A statistical method based on remote sensing for the estimation of air temperature in China. *Intl Journal of Climatology* **2015**, *35*, 2131–2143, doi:10.1002/joc.4113.

3. Golkar, F.; Sabziparvar, A.A.; Khanbilvardi, R.; Nazemosadat, M.J.; Zand-Parsa, S.; Rezaei, Y. Estimation of instantaneous air temperature using remote sensing data. *International Journal of Remote Sensing* **2018**, *39*, 258–275, doi:10.1080/01431161.2017.1382743.

4. Liu, S.; Su, H.; Tian, J.; Zhang, R.; Wang, W.; Wu, Y. Evaluating Four Remote Sensing Methods for Estimating Surface Air Temperature on a Regional Scale. *Journal of Applied Meteorology and Climatology* **2017**, *56*, 803–814, doi:10.1175/JAMC-D-16-0188.1.

5. Jang, K.; Kang, S.; Kimball, J.; Hong, S. Retrievals of All-Weather Daily Air Temperature Using MODIS and AMSR-E Data. *Remote Sensing* **2014**, *6*, 8387–8404, doi:10.3390/rs6098387.

6. Zhu, W.; Lű, A.; Jia, S. Estimation of daily maximum and minimum air temperature using MODIS land surface temperature products. *Remote Sensing of Environment* **2013**, *130*, 62–73, doi:10.1016/j.rse.2012.10.034.

7. Pape, R.; Löffler, J. Modelling spatio-temporal near-surface temperature variation in high mountain landscapes. *Ecological Modelling* **2004**, *178*, 483–501, doi:10.1016/j.ecolmodel.2004.02.019.

8. Hou, P.; Chen, Y.; Qiao, W.; Cao, G.; Jiang, W.; Li, J. Near-surface air temperature retrieval from satellite images and influence by wetlands in urban region. *Theor Appl Climatol* **2013**, *111*, 109–118, doi:10.1007/s00704-012-0629-7.

9. Shen, H.; Jiang, Y.; Li, T.; Cheng, Q.; Zeng, C.; Zhang, L. Deep learning-based air temperature mapping by fusing remote sensing, station, simulation and socioeconomic data. *Remote Sensing of Environment* **2020**, *240*, 111692, doi:10.1016/j.rse.2020.111692.

10. Hooker, J.; Duveiller, G.; Cescatti, A. A global dataset of air temperature derived from satellite remote sensing and weather stations. *Sci. Data* **2018**, *5*, 180246, doi:10.1038/sdata.2018.246.

11. Şahin, M. Modelling of air temperature using remote sensing and artificial neural network in Turkey. *Advances in Space Research* **2012**, *50*, 973–985, doi:10.1016/j.asr.2012.06.021.

12. Zhang, Z.; Du, Q. A Bayesian Kriging Regression Method to Estimate Air Temperature Using Remote Sensing Data. *Remote Sensing* **2019**, *11*, 767, doi:10.3390/rs11070767.

13. Hadria, R.; Benabdelouahab, T.; Mahyou, H.; Balaghi, R.; Bydekerke, L.; El Hairech, T.; Ceccato, P. Relationships between the three components of air temperature and remotely sensed land surface temperature of agricultural areas in Morocco. *International Journal of Remote Sensing* **2018**, *39*, 356–373, doi:10.1080/01431161.2017.1385108.

14. Li, L.; Zha, Y. Estimating monthly average temperature by remote sensing in China. *Advances in Space Research* **2019**, *63*, 2345–2357, doi:10.1016/j.asr.2018.12.039.

15. Liu, S.; Su, H.; Zhang, R.; Tian, J.; Wang, W. Estimating the Surface Air Temperature by Remote Sensing in Northwest China Using an Improved Advection-Energy Balance for Air Temperature Model. *Advances in Meteorology* **2016**, *2016*, 1–11, doi:10.1155/2016/4294219.

16. Xu, Y.; Knudby, A.; Ho, H.C. Estimating daily maximum air temperature from MODIS in British Columbia, Canada. *International Journal of Remote Sensing* **2014**, *35*, 8108–8121, doi:10.1080/01431161.2014.978957.

17. Kloog, I.; Nordio, F.; Coull, B.A.; Schwartz, J. Predicting spatiotemporal mean air temperature using MODIS satellite surface temperature measurements across the Northeastern USA. *Remote Sensing of Environment* **2014**, *150*, 132–139, doi:10.1016/j.rse.2014.04.024.

18. Kim, D.-Y.; Han, K.-S. Remotely sensed retrieval of midday air temperature considering atmospheric and surface moisture conditions. *International Journal of Remote Sensing* **2013**, *34*, 247–263, doi:10.1080/01431161.2012.712235.

19. Gholamnia, M.; Alavipanah, S.K.; Darvishi Boloorani, A.; Hamzeh, S.; Kiavarz, M. Diurnal Air Temperature Modeling Based on the Land Surface Temperature. *Remote Sensing* **2017**, *9*, 915, doi:10.3390/rs9090915.

20. Meyer, H.; Katurji, M.; Appelhans, T.; Müller, M.; Nauss, T.; Roudier, P.; Zawar-Reza, P. Mapping Daily Air Temperature for Antarctica Based on MODIS LST. *Remote Sensing* **2016**, *8*, 732, doi:10.3390/rs8090732.

21. Benali, A.; Carvalho, A.C.; Nunes, J.P.; Carvalhais, N.; Santos, A. Estimating air surface temperature in Portugal using MODIS LST data. *Remote Sensing of Environment* **2012**, *124*, 108–121, doi:10.1016/j.rse.2012.04.024.

22. Samanta, S.; Pal, D.K.; Lohar, D.; Pal, B. Interpolation of climate variables and temperature modeling. *Theor Appl Climatol* **2012**, *107*, 35–45, doi:10.1007/s00704-011-0455-3.

23. Wang, M.; He, G.; Zhang, Z.; Wang, G.; Zhang, Z.; Cao, X.; Wu, Z.; Liu, X. Comparison of Spatial Interpolation and Regression Analysis Models for an Estimation of Monthly Near Surface Air Temperature in China. *Remote Sensing* **2017**, *9*, 1278, doi:10.3390/rs9121278.

24. Nikoloudakis, N.; Stagakis, S.; Mitraka, Z.; Kamarianakis, Y.; Chrysoulakis, N. Spatial interpolation of urban air temperatures using satellite-derived predictors. *Theor Appl Climatol* **2020**, *141*, 657–672, doi:10.1007/s00704-020-03230-3.

25. Ruiz-Álvarez, M.; Alonso-Sarria, F.; Gomariz-Castillo, F. Interpolation of Instantaneous Air Temperature Using Geographical and MODIS Derived Variables with Machine Learning Techniques. *IJGI* **2019**, *8*, 382, doi:10.3390/ijgi8090382.

26. Fries, A.; Rollenbeck, R.; Göttlicher, D.; Nauß, T.; Homeier, J.; Peters, T.; Bendix, J. Thermal structure of a megadiverse Andean mountain ecosystem in southern Ecuador and its regionalization. *erd* **2009**, *63*, 321–335, doi:10.3112/erdkunde.2009.04.03.

27. Jing, W.; Yang, Y.; Yue, X.; Zhao, X. A Comparison of Different Regression Algorithms for Downscaling Monthly Satellite-Based Precipitation over North China. *Remote Sensing* **2016**, *8*, 835, doi:10.3390/rs8100835.

28. Kang, L.; Di, L.; Deng, M.; Shao, Y.; Yu, G.; Shrestha, R. Use of Geographically Weighted Regression Model for Exploring Spatial Patterns and Local Factors Behind NDVI-Precipitation Correlation. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing* **2014**, *7*, 4530–4538, doi:10.1109/JSTARS.2014.2361128.

29. Shen, J.; Liu, P.; Xia, J.; Zhao, Y.; Dong, Y. Merging Multisatellite and Gauge Precipitation Based on Geographically Weighted Regression and Long Short-Term Memory Network. *Remote Sensing* **2022**, *14*, 3939, doi:10.3390/rs14163939.

30. Hu, D.; Shu, H.; Hu, H.; Xu, J. Spatiotemporal regression Kriging to predict precipitation using time-series MODIS data. *Cluster Comput* **2017**, *20*, 347–357, doi:10.1007/s10586-016-0708-0.

31. Bostan, P.A.; Heuvelink, G.B.M.; Akyurek, S.Z. Comparison of regression and kriging techniques for mapping the average annual precipitation of Turkey. *International Journal of Applied Earth Observation and Geoinformation* **2012**, *19*, 115–126, doi:10.1016/j.jag.2012.04.010.

32. Moraux, A.; Dewitte, S.; Cornelis, B.; Munteanu, A. A Deep Learning Multimodal Method for Precipitation Estimation. *Remote Sensing* **2021**, *13*, 3278, doi:10.3390/rs13163278.

33. Bajat, B.; Pejović, M.; Luković, J.; Manojlović, P.; Ducić, V.; Mustafić, S. Mapping average annual precipitation in Serbia (1961–1990) by using regression kriging. *Theor Appl Climatol* **2013**, *112*, 1–13, doi:10.1007/s00704-012-0702-2.

34. Lu, X.; Li, J.; Liu, Y.; Li, Y.; Huo, H. Quantitative Precipitation Estimation in the Tianshan Mountains Based on Machine Learning. *Remote Sensing* **2023**, *15*, 3962, doi:10.3390/rs15163962.

35. Varouchakis, E.A.; Kamińska-Chuchmała, A.; Kowalik, G.; Spanoudaki, K.; Graña, M. Combining Geostatistics and Remote Sensing Data to Improve Spatiotemporal Analysis of Precipitation. *Sensors (Basel)* **2021**, *21*, doi:10.3390/s21093132.

36. Hurter, F.; Maier, O. Tropospheric profiles of wet refractivity and humidity from the combination of remote sensing data sets and measurements on the ground. *Atmos. Meas. Tech.* **2013**, *6*, 3083–3098, doi:10.5194/amt-6-3083-2013.

37. Yan, D.; Liu, T.; Dong, W.; Liao, X.; Luo, S.; Wu, K.; Zhu, X.; Zheng, Z.; Wen, X. Integrating remote sensing data with WRF model for improved 2-m

temperature and humidity simulations in China. *Dynamics of Atmospheres and Oceans* **2020**, *89*, 101127, doi:10.1016/j.dynatmoce.2019.101127.

38. Cai, X.; Bao, Y.; Petropoulos, G.P.; Lu, F.; Lu, Q.; Zhu, L.; Wu, Y. Temperature and Humidity Profile Retrieval from FY4-GIIRS Hyperspectral Data Using Artificial Neural Networks. *Remote Sensing* **2020**, *12*, 1872, doi:10.3390/rs12111872.

39. Che, Y.; Ma, S.; Xing, F.; Li, S.; Dai, Y. An improvement of the retrieval of temperature and relative humidity profiles from a combination of active and passive remote sensing. *Meteorol Atmos Phys* **2019**, *131*, 681–695, doi:10.1007/s00703-018-0588-3.

40. Jiang, J.H.; Yue, Q.; Su, H.; Kangaslahti, P.; Lebsock, M.; Reising, S.; Schoeberl, M.; Wu, L.; Herman, R.L. Simulation of Remote Sensing of Clouds and Humidity From Space Using a Combined Platform of Radar and Multifrequency Microwave Radiometers. *Earth and Space Science* **2019**, *6*, 1234–1243, doi:10.1029/2019EA000580.

41. Jackson, D.L.; Wick, G.A.; Bates, J.J. Near-surface retrieval of air temperature and specific humidity using multisensor microwave satellite observations. *J. Geophys. Res.* **2006**, *111*, 755, doi:10.1029/2005JD006431.

42. Polyakov, A.; Virolainen, Y.; Nerobelov, G.; Timofeyev, Y.; Solomatnikova, A. Total ozone measurements using IKFS-2 spectrometer aboard Meteor-M N2 satellite in 2019–2020. *International Journal of Remote Sensing* **2021**, *42*, 8709–8733, doi:10.1080/01431161.2021.1985741.

43. Şenkal, O. Modeling of solar radiation using remote sensing and artificial neural network in Turkey. *Energy* **2010**, *35*, 4795–4801, doi:10.1016/j.energy.2010.09.009.

44. Robles-Zazueta, C.A.; Molero, G.; Pinto, F.; Foulkes, M.J.; Reynolds, M.P.; Murchie, E.H. Field-based remote sensing models predict radiation use efficiency in wheat. *J. Exp. Bot.* **2021**, *72*, 3756–3773, doi:10.1093/jxb/erab115.

45. Ghimire, S.; Deo, R.C.; Downs, N.J.; Raj, N. Self-adaptive differential evolutionary extreme learning machines for long-term solar radiation prediction with remotely-sensed MODIS satellite and Reanalysis atmospheric products in solar-rich cities. *Remote Sensing of Environment* **2018**, *212*, 176–198, doi:10.1016/j.rse.2018.05.003.

46. Ghimire, S.; Deo, R.C.; Downs, N.J.; Raj, N. Global solar radiation prediction by ANN integrated with European Centre for medium range weather forecast fields in solar rich cities of Queensland Australia. *Journal of Cleaner Production* **2019**, *216*, 288–310, doi:10.1016/j.jclepro.2019.01.158.

47. Yan, G.; Wang, T.; Jiao, Z.; Mu, X.; Zhao, J.; Chen, L. Topographic radiation modeling and spatial scaling of clear-sky land surface longwave radiation over rugged terrain. *Remote Sensing of Environment* **2016**, *172*, 15–27, doi:10.1016/j.rse.2015.10.026.

48. Samani, Z.; Bawazir, A.S.; Bleiweiss, M.; Skaggs, R.; Tran, V.D. Estimating Daily Net Radiation over Vegetation Canopy through Remote Sensing and Climatic Data. *J. Irrig. Drain Eng.* **2007**, *133*, 291–297, doi:10.1061/(ASCE)0733-9437(2007)133:4(291).

49. Whitlock, C.H.; Charlock, T.P.; Staylor, W.F.; Pinker, R.T.; Laszlo, I.; Ohmura, A.; Gilgen, H.; Konzelman, T.; DiPasquale, R.C.; Moats, C.D.; et al. First Global WCRP Shortwave Surface Radiation Budget Dataset. *Bull. Amer. Meteor. Soc.* **1995**, *76*, 905–922, doi:10.1175/1520-0477(1995)076<0905:FGWSSR>2.0.CO;2.