SOFTWARE TOOL ARTICLE

REVISED Scalable, open-access and multidisciplinary data integration pipeline for climate-sensitive diseases

[version 2; peer review: 2 approved]

Abhishek Dasgupta[1,2], Iago Perez-Fernandez[3], Tuyen Huynh[4], Cathal Mills[2,5], Rowan C. Nicholls[1,2], Prathyush Sambaturu[2,6], Marc Choisy[4,7], David Wallom[3], Tung Nguyen-Duy[4], Rhys P. D. Inward[2,6], John-Stuart Brittain[1,2], Sarah Sparrow[3], Moritz U.G. Kraemer[2,6]

[1]Oxford Research Software Engineering Group, Doctoral Training Centre, University of Oxford, Oxford, UK
[2]Pandemic Sciences Institute, University of Oxford, Oxford, England, UK
[3]Oxford e-Research Centre, Department of Engineering Science, University of Oxford, Oxford, England, UK
[4]Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam
[5]Department of Statistics, University of Oxford, Oxford, England, UK
[6]Department of Biology, University of Oxford, Oxford, UK
[7]Nuffield Department of Medicine, University of Oxford, Oxford, England, UK

v2 **First published:** 29 Aug 2025, **10**:467
https://doi.org/10.12688/wellcomeopenres.24774.1
Latest published: 15 Nov 2025, **10**:467
https://doi.org/10.12688/wellcomeopenres.24774.2

Abstract

Climate-sensitive infectious diseases pose an important challenge for human, animal and environmental health and it has been estimated that over half of known human pathogenic diseases can be aggravated by climate change. While climatic and weather conditions are important drivers of transmission of vector-borne diseases, socio-economic, behavioural, and land-use factors as well as the interactions among them impact transmission dynamics. Analysis of drivers of climate-sensitive diseases require rapid integration of interdisciplinary data to be jointly analysed with epidemiological (including genomic and clinical) data. Current tools for the integration of multiple data sources are often limited to one data type or rely on proprietary data and software. To address this gap, we develop a scalable and open-access pipeline for the integration of multiple spatio-temporal datasets that requires only the declaration of the country and temporal range and resolution of the study. The tool is locally deployable and can easily be integrated into existing climate-disease-modelling applications. We demonstrate the utility of the tool for dengue modelling in Vietnam where epidemiological data are legally required to remain local. We include a pipeline for bias correction of climate data to enhance their quality for downstream

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 15 Nov 2025	✓ view	✓ view
version 1 29 Aug 2025	✗ view	✓ view

1. **Oliver Brady**, London School of Hygiene & Tropical Medicine, London, UK

2. **Kayode Oshinubi**, Northern Arizona University, Flagstaff, USA

Any reports and responses or comments on the article can be found at the end of the article.

modelling tasks. The Dengue Advanced Readiness Tools-Pipeline empowers users by simplifying complex download, correction, and aggregation steps, fostering data-driven discovery of relationships between infectious diseases and their drivers in space and time, and enhancing reproducibility in research. Additional modules and datasets can be added to the existing ones to make the pipeline extendable to use cases other than the ones presented here.

Plain language summary

Most human infectious diseases are affected by climate. Therefore, understanding how climate and other rapidly changing factors like land use, population behaviour, and socio-economic conditions influence the spread of these diseases requires combining many different types of data. However, tools that enable researchers and policy makers to bring together and use this data remain limited, difficult to use, or rely on software that is not freely available.

To address this gap, we developed a new, open-access software pipeline that helps users collect, clean, and prepare different types of data—such as weather, socio-economic, and health data—for analysis. The pipeline works with data across different locations and time periods, and only requires users to specify the country and timeframe they are interested in. It is designed to be easy to use and can be run on local computers, which is especially important in places where health data is not allowed to leave the country. We deployed the pipeline using data on dengue fever in Vietnam. By making the process simpler and more transparent, we hope to support faster and more reliable responses to climate-related health threats.

Keywords

data science, automated workflows, climate-sensitive infectious diseases, dengue

Corresponding authors: Abhishek Dasgupta (abhishek.dasgupta@dtc.ox.ac.uk), Moritz U.G. Kraemer (moritz.kraemer@biology.ox.ac.uk)

Author roles: Dasgupta A: Conceptualization, Formal Analysis, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Perez-Fernandez I**: Formal Analysis, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Huynh T**: Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Mills C**: Writing – Original Draft Preparation, Writing – Review & Editing; **Nicholls RC**: Conceptualization, Formal Analysis, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Sambaturu P**: Conceptualization, Formal Analysis, Writing – Original Draft Preparation, Writing – Review & Editing; **Choisy M**: Writing – Original Draft Preparation, Writing – Review & Editing; **Wallom D**: Writing – Original Draft Preparation, Writing – Review & Editing; **Nguyen-Duy T**: Writing – Original Draft Preparation, Writing – Review & Editing; **Inward RPD**: Formal Analysis, Writing – Original Draft Preparation, Writing – Review & Editing; **Brittain JS**: Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Sparrow S**: Conceptualization, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; **Kraemer MUG**: Conceptualization, Formal Analysis, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome [303666; 226052; 228186 to M.U.G.K]; This tool was developed as part of the Global.health project (https://global.health/). M.U.G.K. acknowledges funding from The Rockefeller Foundation [PC-2022-POP-005], Google.org, the Oxford Martin School Programmes in Pandemic Genomics & Digital Pandemic Preparedness, European Union's Horizon Europe programme projects MOOD [#874850] and E4Warning [#101086640], the United Kingdom Research and Innovation [#APP8583], the Medical Research Foundation [MRF-RG-ICCH-2022-100069], UK International Development [301542-403], the Bill & Melinda Gates Foundation [INV-063472] and Novo Nordisk Foundation [NNF24OC0094346]. T.H. is supported by the Nuffield Department of Medicine Tropical Network Fund. The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission or the other funders.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Dasgupta A, Perez-Fernandez I, Huynh T *et al.* **Scalable, open-access and multidisciplinary data integration pipeline for climate-sensitive diseases [version 2; peer review: 2 approved]** Wellcome Open Research 2025, **10**:467 https://doi.org/10.12688/wellcomeopenres.24774.2

First published: 29 Aug 2025, **10**:467 https://doi.org/10.12688/wellcomeopenres.24774.1

REVISED **Amendments from Version 1**

In the revised manuscript, we have clarified that while the current pipeline does not support processing epidemiological data that is widely varied and can require bespoke processing pipelines for data imputation and cleaning, our pipeline is modular and can be extended to support such functionality in the future. Figure 1 has been updated to show that the model is user-supplied and not included in the pipeline.

We have added information on non-raster, tabular data that is supported by the pipeline, such as Meta's Relative Wealth Index, and noted that our pipeline supports custom shapefiles. We have also illustrated the end-to-end operation of the pipeline using synthetic dengue case incidence data.

Any further responses from the reviewers can be found at the end of the article

Introduction

Climate-sensitive infectious diseases represent a growing public health challenge with changes in weather and climate that can create conditions that influence disease transmission[1–4]. Recent estimates suggest that over half of known human pathogenic diseases may be aggravated by climate change, underscoring the urgent need for enhanced understanding of the mechanisms linking environmental change and disease dynamics[4]. While climatic and weather variables such as temperature, precipitation, and humidity are well-recognised drivers of vector-borne diseases and respiratory viruses[5], non-climatic factors including demographic, socio-economic conditions, human behavior, and land-use changes also play critical roles[6–9]. Understanding how these diverse drivers interact requires an integrative approach that bridges multiple scientific disciplines and data types ranging from tabular to satellite imagery.

Despite growing recognition of the need for interdisciplinary approaches, existing tools for integrating spatio-temporal data relevant to climate-sensitive diseases remain limited[9,10]. Many platforms focus on specific data types, such as Google Earth Engine for satellite imagery (https://earthengine.google.com/), or rely on proprietary datasets and software, such as Atlas AI (https://www.atlasai.co/). This fragmentation prohibits comprehensive analyses by requiring researchers to manage disparate platforms and potentially apply complex and time-consuming preprocessing steps manually[10]. These challenges are further compounded when working with epidemiological data that may be sensitive or legally constrained from leaving local contexts requiring flexibility for tools to be easily locally deployable.

To address these limitations, we developed the DART (Dengue Advanced Readiness Tools, https://www.dartdengue.org/) pipeline, an open-access, locally deployable system designed to streamline the integration of diverse spatio-temporal epidemiological, socio-economic, and climatic datasets. After defining the spatial extent of a study area (country ISO3 code) with the desired administrative unit and the time frame of the study, researchers can automatically access, download, and preprocess environmental, and socio-economic data for joint downstream analyses at daily, weekly or monthly timesteps. The pipeline supports data integration of epidemiological data, enabling incorporation into climate-disease modelling applications and more broadly real-time modelling tasks.

We demonstrate the capabilities of the DART pipeline through a case study on dengue transmission modelling in Vietnam, where epidemiological data must remain locally stored. Comprehensive documentation and user guides are available at: https://dart-pipeline.readthedocs.io/en/latest/.

Methods

The variables included in our pipeline were selected primarily for forecasting dengue cases[11,12], based on their previously estimated impacts on disease transmission, influencing and modulating vector and pathogen survival, development, and reproduction, and human susceptibility and exposure[13–17]. These variables have also been used in previous forecasting studies for climate-sensitive diseases other than dengue (e.g. West Nile Virus[18]). Table 1 shows the current challenges for robust analyses of climate-sensitive diseases and the functionality of the proposed digital tool. Figure 1 shows the workflow involved in the pipeline from initial download to final storage of files in netCDF format.

Observational data: In this study we used reanalysis data to represent the weather conditions observed in a specific period of time. Reanalysis data shows an accurate estimate of the state of the atmosphere/ocean for a particular time step, and this dataset is obtained by combining observations and short range weather forecasts (1–3 days in advance) computed by numerical weather prediction models[19,20]. Even though reanalysis data shows a very accurate representation of the weather conditions, they might deviate from the observations for some regions[19], affecting the forecasting skill of the Dengue prediction model. Hence applying a bias correction technique will reduce the deviation between observations and predictions. Bias correction techniques are applied to correct precipitation data using a method called quantile mapping[21,22]. This technique is based on the interpolation of the quantile points from a target dataset to a reference dataset so the corrected distribution has statistical properties closer to the observations. Here, we used Vietnam gridded precipitation data[23] as the reference and ERA5[20] and applied the correction following a daily timestep. We also replace outliers above 99th percentile in the reference data by the corresponding 99th percentile value as quantile mapping struggles at correcting extremely high values of daily precipitation, inflating extreme precipitation values obtained in the correction[21]. This way we avoid the manifestation of anomalously high precipitation values in the correction.

Socio-economic data. In addition to observational data, we integrated Meta's relative wealth index (RWI) data into the pipeline following the processing steps outlined in (https://dataforgood.facebook.com/dfg/docs/tutorial-calculating-population-weighted-relative-wealth-index). The RWI data are provided by Meta as tabular point-level estimates indexed by quadkeys,

Table 1. Challenges in the generation and execution of multidisciplinary data integration pipelines in order of consideration.

Process	Challenges	*DART* pipeline functionality
Dataset identification	• Availability of a large number of covariates, including multiple for the same variables • Requires domain expertise from multiple different domains (epidemiological, climatic, environmental, socio-economic)	• Curated set of covariates that work well together • Domain expertise is not needed as all variables are aggregated to a unified schema for easy comparison
Data download	• No automated download functionality across different datasets and domains • Issues with data licensing and attribution	• Single interface to download data across different domains and process to an unified schema • With a few exceptions, curation from public data sources and documented attribution
Data aggregation	• Raw data are rarely at the temporal and spatial scales needed for downstream applications • Aggregation conventions differ across domains requiring human input, capabilities and judgement • Bespoke code needed for aggregating spatio-temporal data	• Pre-processing steps in the pipeline ensure appropriate temporal and spatial resolutions, with aggregation to daily and weekly levels in temporal scale • Aggregation of data to administrative units using appropriate aggregation methodology for each variable. Pre-selected aggregation methods that can be extended or replaced • In addition to standard aggregation functions (sum, mean), population weighting of meteorological variables is performed as it is often more relevant for infectious disease research
Data validation	• Large data pipelines are often error-prone which could impact downstream applications	• Unit and integration testing, along with visual checks by experts
Data integration	• Integrating data across epidemiological, climate, environmental, and socio-economic domains is challenging due to differences in spatial and temporal definitions, references, coverage and resolutions • Epidemiological data are often sensitive and cannot be shared outside of specific agencies	• Unified data storage in netCDF with (time, region) coordinates and annotated with CF-compliant and DART specific metadata that allows comparison to private datasets that cannot be included within the public pipeline

each representing a ~2.4 km grid cell. Each record includes latitude and longitude coordinates corresponding to the grid cell centroid, enabling spatial mapping and integration with other geospatial datasets. In our implementation, RWI values are aggregated to administrative units with population-weights (mean), with Meta's population density data serving as weights. Latitude and longitude are used primarily for spatial referencing and visualization, while the quadkey identifier ensures consistent alignment between RWI and population layers during aggregation. The framework also supports replacing Meta's population data with census-derived population grids to calculate population-weighted RWI or incorporate alternative socioeconomic indicators.

Weather forecasting: In addition to using reanalysis data, we also included future weather forecast data up to two weeks ahead. Weather forecast data is obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF), specifically from the ensemble prediction system (hereafter ECMWF weather forecast model). ECMWF weather forecast model is a probabilistic prediction model that estimates the most probable state of the atmosphere in the future while quantifying the uncertainty of the predictions up to 15 days in advance. Nonetheless, the further we go into the future, the more uncertain the forecast becomes, hence weather forecast data is often post-processed to correct possible biases so it can be more useful for applications. Prior studies have shown that the weather variables that are most linked to Dengue incidence are temperature, relative humidity and accumulated precipitation, and early results from DART showed that applying quantile mapping to raw weather forecast data increases the reliability and accuracy of the predictions beyond 10 days[22,24]. We also applied quantile mapping to calibrate ECMWF weather forecast data, using ERA5 data as reference. The correction is applied separately to the weather forecast 11 ensemble members (10 perturbed + 1 control).

In the following couple of figures, we show the effect of bias correction on weather variables in a region around Ho Chi Minh City, Vietnam (HCMC), which is our area of interest for the

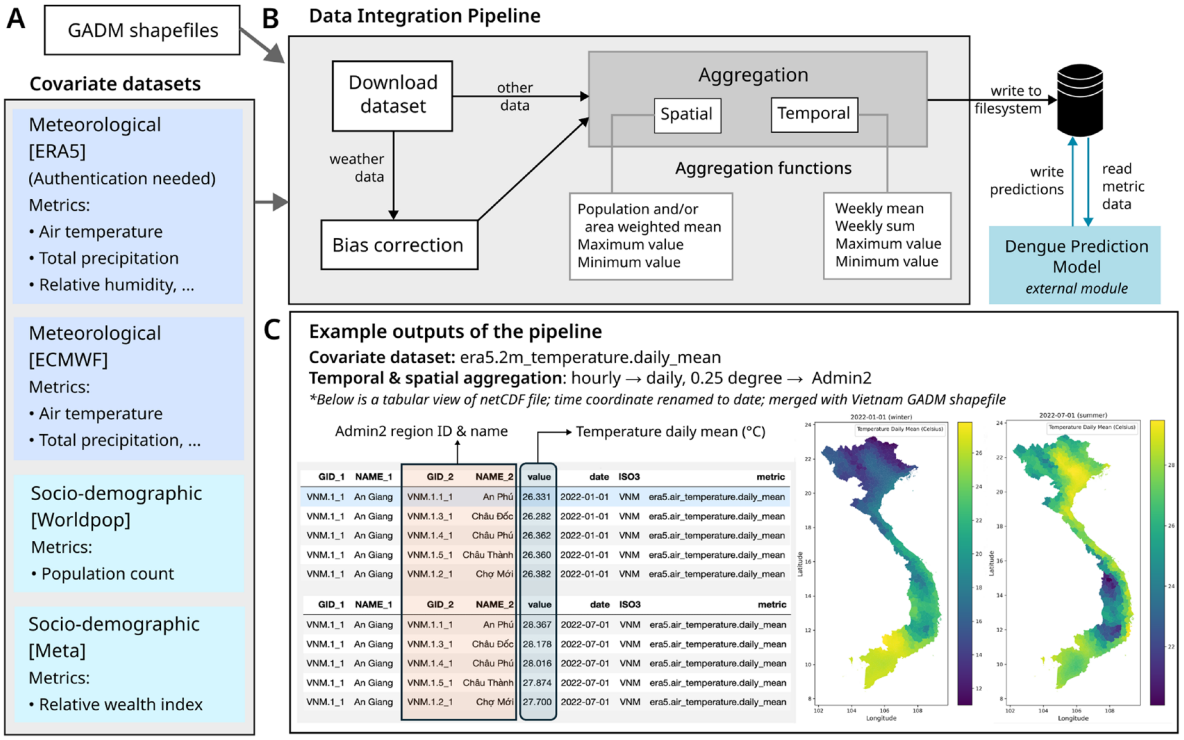

Figure 1. Workflow of the Data Integration Pipeline and Example Outputs. A) Overview of the input datasets used in the integration pipeline. These datasets include various covariates associated with dengue incidence, categorised into meteorological and socio-demographic factors. For each dataset, a selection of relevant variables, data sources, and access restrictions, where applicable, is provided. Additionally, shapefiles from the Database of Global Administrative Areas (GADM) were used to define the geographical boundaries of administrative units at multiple levels (e.g., admin2 and admin3) considered in the pipeline. **B**) The integration pipeline consists of multiple sequential steps. First, datasets containing the relevant variables are downloaded from their respective sources. A dedicated sub-pipeline corrects weather data bias using quantile mapping technique. Finally, spatial and temporal aggregations are performed at the users predefined administrative levels. Spatial aggregation methods—such as population-weighted or area-weighted means, maximum, and minimum values—are applied based on the nature of each variable, using GADM shapefiles for administrative levels admin2 and weighted using Worldpop population data. Similarly, temporal aggregation methods, including weekly mean, sum, maximum, and minimum values, are computed. The processed data is then stored in a structured schema in the standard netCDF file format, which includes region ID, datetime (corresponding to the variable's weekly value), variable metadata, and the aggregated value. Further, the aggregated dataset is incorporated into dengue modelling for further analysis, and the model's outputs are subsequently stored in the database. **C**) This panel illustrates example outputs using the air temperature at 2m variable from the ERA5 dataset. The data is temporally aggregated (see Table S1 for methods) and then spatially aggregated to the admin2 level. The tables display the daily mean air temperature for admin2 regions on two specific dates: January 1, 2022, and July 7, 2022. Corresponding maps on the right visualise this data, where lighter shades represent higher daily mean temperatures, and darker shades indicate lower temperatures, providing the spatial distribution of temperature variations.

pipeline implementation. Figure 2 shows the mean difference (or bias) between raw (Figure 2a–c) and corrected ECMWF weather forecast model estimates (Figure 2d–f) against ERA5 for 2 metre temperature, relative humidity and total precipitation between 1–2 weeks in advance. Raw ECMWF forecast model predictions show a systematic cold bias, underestimating 2 metre temperature by ~2 degrees (Figure 2a). Mean relative humidity does not appear to vary much from observations (it just shows a small positive bias) (Figure 2b). In the case of accumulated precipitation (Figure 2c), there is a positive bias (the model overestimates precipitation) around 20 mm in the coastal/oceanic regions and slightly less in HCMC city. After applying the quantile mapping technique, the biases between observations and corrected forecasts (Figure 2d–f), are practically negligible.

Another example of the performance of the calibration can be seen in Figure 3, where we show the anomaly correlation coefficient (ACC), between raw (Figure 3a–c) and corrected forecast data (Figure 3d,f). ACC quantifies the ability of the forecast model to predict values that deviate from the climatological mean. These climate anomalies should be accurately modelled as it is especially important for climate-sensitive diseases such as dengue. A recent study[25] suggested extreme weather events can lead to increased dengue risk, and another[26] identified how climate anomalies can be used to predict long-term trends of dengue. When ACC >0.6 the weather forecast is considered skillful. Here, we used ERA5 as reference data; raw ECMWF data for the first lead week of prediction shows ACC values around 0.8–0.9 for 2 metre temperature for the first week, and manages to remain skillful to the 2nd week near HCMC

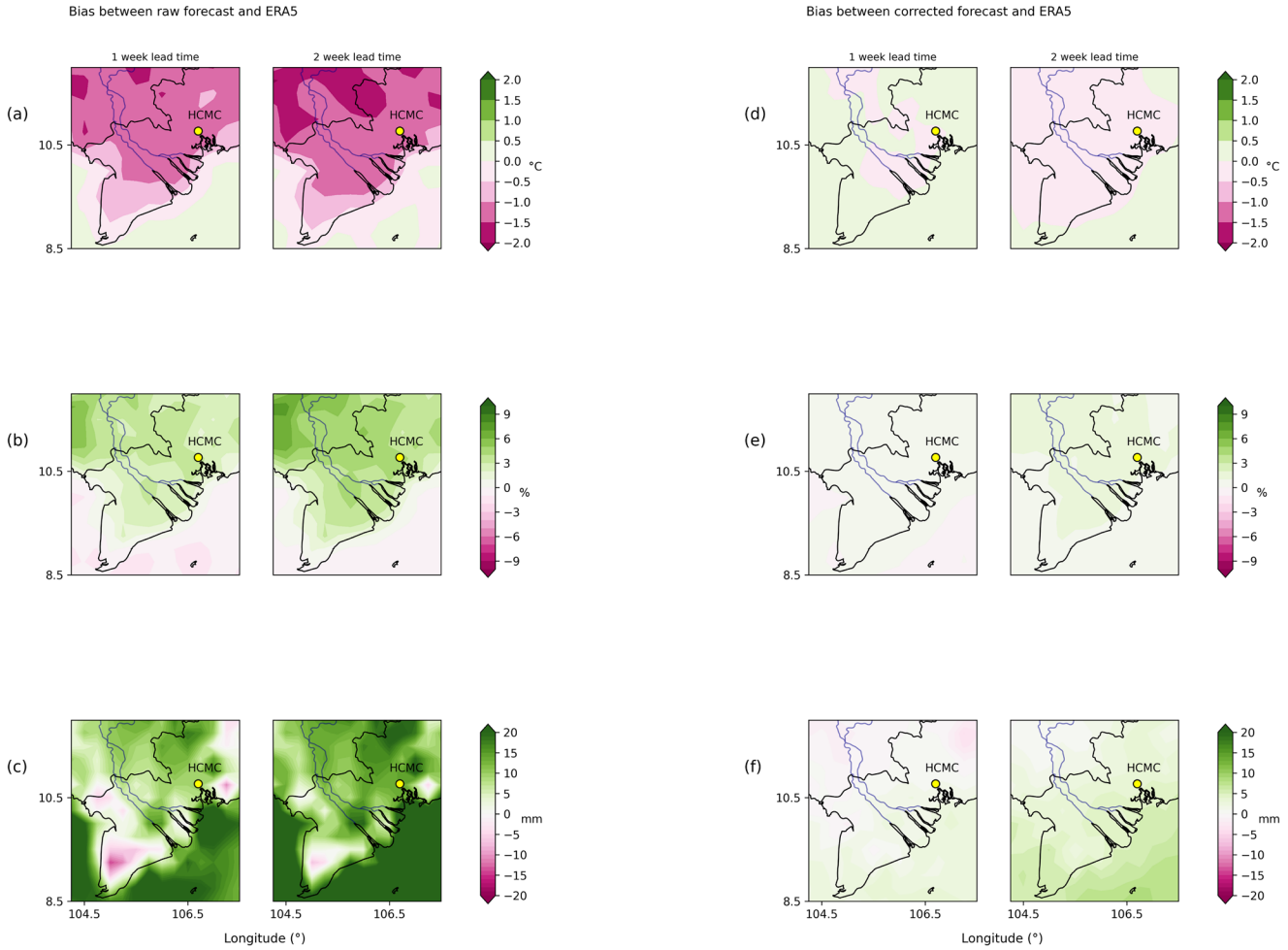

Figure 2. Bias between raw and corrected ECMWF weather forecast model against ERA5 in South Vietnam for 2 metre temperature (Figure **2a**,**d**), relative humidity (Figure **2b**,**e**) and total precipitation (Figure **2c**,**f**) between 1–2 weeks in advance between 2004–2020. Figure **2a** (**2d**) shows results for raw (corrected) 2 metre temperature, **2b** (**2e**) for raw (corrected) relative humidity and **2c** (**2f**) for raw (corrected) accumulated precipitation.

(Figure 3a). In the case of relative humidity (Figure 3b), results are similar to 2 metre temperature, with the difference that by the 2nd week the forecast is barely skillful. By contrast, when we measure the ACC using total precipitation, we find the lowest values during the 1st week of forecast (between 0.6–0.7 and in some areas we have no skill), and there is practically no skill at the 2nd forecast week (Figure 3c). After applying the quantile mapping technique (Figure 3d–f), we generally obtain higher values of ACC for the first week for every variable, (ACC values between 0.8–1.0). Forecast skill is improved 2 weeks ahead for all variables in South Vietnam, although for relative humidity (Figure 3e) the skill is slightly lower compared to the other variables, but still well above the threshold (ACC>0.6).

Results from Figure 2 and Figure 3 show that the quantile mapping technique extends the reliability and accuracy of the forecast up to 2 weeks in advance, hence we applied this technique to correct 2 metre temperature, relative humidity and total precipitation for the weather forecast in order to obtain more accurate predictions of these variables.

ECMWF only makes ERA5 data available with a lag of 5 days which results in a 5-day gap between observations and the current date. In order to fill this 5 day gap critical for decision making, we obtain the weather forecast that initialises 7 days before the current date (t) and retain data that goes up to day t+7. We then perform bias correction on this weather forecast, aggregate daily forecasts into a one week prediction, and use this as a "pseudo-observation", inputting it to the dengue prediction model for the following week.

Spatial aggregation: The climatic weather variables are stored in a spatio-temporal raster, while administrative unit boundaries are polygons (see Figure S1). We performed temporal and spatial aggregation to the administrative level

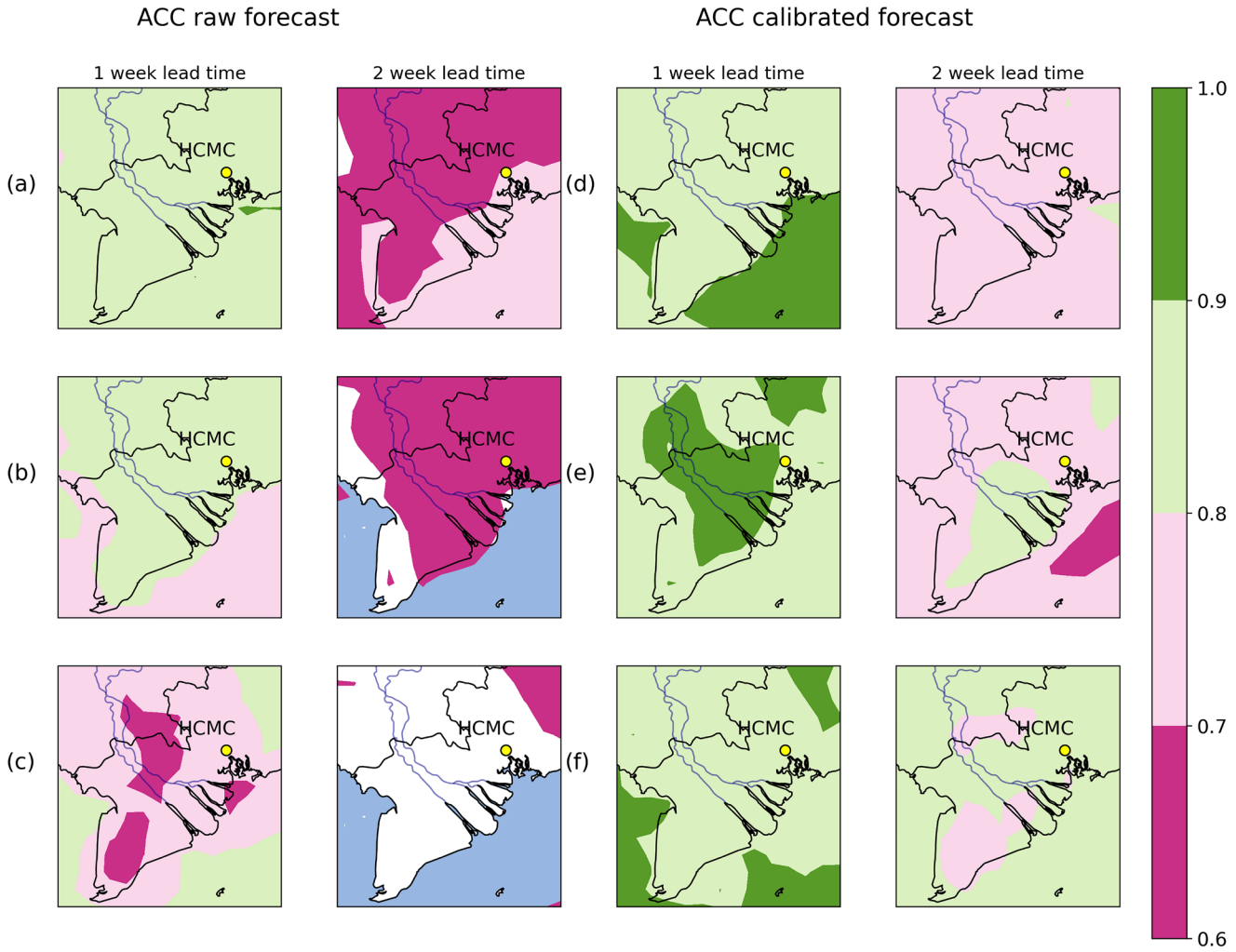

Figure 3. Anomaly correlation coefficient (ACC) for the region of South Vietnam for lead weeks 1–2 for uncalibrated (**a**–**c**) and calibrated (**d**–**f**) ECMWF weather forecast model and ERA5 data between 2008–2024. Figure **3a** (**3d**) shows results for raw (corrected) 2 metre temperature, **3b** (**3e**) for raw (corrected) relative humidity and **3c** (**3f**) for raw (corrected) accumulated precipitation.

that matches the resolution of the epidemiological data. Due to heterogeneity in population distributions within administrative units, we performed population weighting at 1km x 1km resolution. This aggregation process was performed using WorldPop population count rasters, ERA5 weather rasters and administrative boundaries from GADM[27]. Users can optionally provide their own shapefiles via a custom configuration. The process follows the following six steps:

(i) All rasters are projected onto EPSG:4326 (standard latitude/longitude, WGS84 datum)

(ii) All raster files are cropped to the extent of the administrative units of interest (regions can be specified by the user)

(iii) Rasters are resampled to match the resolution of the population raster (in our case 1km x 1km but can be user defined with a custom function)

(iv) Each raster cell's weighting is calculated according to a formula that includes the coverage fraction of each raster cell with the administrative boundary polygon from the shapefile, with weights proportional to the spherical area in square kilometres.

(v) Each raster cell is weighted by the population cell value, normalised by each polygons total population

(vi) Cell values are aggregated to the administrative units using either the population weighted mean or area weighted sum for instantaneous and accumulative variables respectively.

Temporal aggregation: Covariates have varying spatial resolutions with climate variables downloaded at hourly timesteps and aggregated to daily or weekly mean, maximum or minimum values per pixel. After the temporal aggregation, we perform spatial aggregation to a user-defined administrative level. We use the ISO-8601 standard for date (https://www.iso.org/iso-8601-date-and-time-format.html) with weeks starting on Mondays.

Data dictionary: All output data are stored as netCDF files with one netCDF file per data source (historical ERA5 weather data, ECMWF forecast data and socio-demographic data). Metadata are stored as netCDF attributes and adhere closely to CF conventions (https://cfconventions.org). Each netCDF has variables that are indicated by CF-compliant or commonly used short names (such as t2m for air temperature at 2m, and r for relative humidity). The following metadata attributes are present. Each variable in a netCDF file is defined by (time, region) coordinates.

Global attributes:

- DART_region: Global netCDF attribute, describes the region for which the zonal statistics was performed. This includes the region or ISO3 code, the shapefile column used to derive coordinates for 'region' (in our case, GID_2, the unique ID for GADM administrative level 2; user specified ID in case of custom configuration), the timezone (used for timeshifting hourly ERA5 data) and geospatial extents as a bounding box.

Coordinates:

- time: time coordinate for aggregated variable value (mean for instantaneous, sum for accumulative variables) until the next time step
- region: unique ID that denotes the administrative region over which zonal aggregation was performed

Variable attributes: These attributes are CF-compliant

- long_name: Description of the variable
- units: Name of the unit, as recommended by CF conventions *udunits* package[28]
- cell_methods: indicating the temporal aggregation steps
- valid_min (optional): minimum valid value
- valid_max (optional): maximum valid value
- standard_name (optional): CF-compliant standard name if one exists

Software Architecture: The pipeline is written in *Python 3* and depends on well-known libraries in the geospatial Python ecosystem such as *rasterio* and *xarray* for raster manipulation, and *geopandas* and *shapely* for shapefile processing. To ensure reproducibility, we use a lock file which pins package versions, preventing package upgrades from breaking the pipeline; we will also attempt to keep the pipeline updated with the latest major versions of underlying dependencies. We chose *Python* for its extensive library support for data science and machine learning. We also provide a command line interface to make our pipeline interoperable with other tools and programming languages. We provide a user-friendly method to download data, with user prompts in cases where authenticated access is required, such as for ERA5 open data from ECMWF. Data are downloaded into standardised locations described in the documentation. Data processing is performed as described in the Methods section, where we use a helper library which we authored and called *geoglue* (available at https://github.com/kraemer-lab/geoglue), that enables higher level geospatial operations in *Python* compared to those offered by *rasterio*. For climate data, resampling to the target resolution of the population raster is performed by the Climate Data Operators library using the *remapbil* operator for bilinear interpolation (for instantaneous variables) and the *remapdis* operator for distance-weighted average remapping (for accumulative variables). Zonal statistics is performed using the *exactextract* package using operations suitable for each variable (weighted_mean for instantaneous, area_weighted_sum for accumulative variables, Table S2). After performing zonal aggregation, we annotate the data with appropriate metadata as described above.

Compared to prior packages for zonal aggregation in *Python*, such as *rasterstats*, the dependency in *geoglue* that we use for zonal aggregation, *exactextract*, can use the overlap fraction of a raster cell with a polygon, rather than entirely including or omitting a raster cell based on partial overlap which gives better results. We compute the weighted mean or area-weighted sum for each administrative region. As an illustration for this manuscript, we demonstrate the pipeline functionality for Vietnam, aggregating to Global Administrative Level 2 (i.e. district-level subdivisions, henceforth called "district", n = 710[27].

Data architecture: We store output data in flat files in the open-source and well-documented binary netCDF file format. We chose netCDF compared to the ASCII text file formats such as CSV due to its widespread use in climate data processing and for its rich metadata support. Libraries such as *xarray* and extensions of netCDF such as *zarr* and *icechunk* also enable future scaling to larger datasets. Storing data in a well-documented binary format along with libraries like *xarray* makes it possible to only work with a portion of the data and allow easy reshaping and resampling of the data. Using flat files enables scaling the solution from a users' laptop to the data center as data size and computational complexity of running the pipeline increases. It also enables simpler parallelization for large source datasets; by writing to a shared filesystem (or network data store), scripts can run in parallel on a laptop or use HPC infrastructure to parallelise fetching and processing data.

Testing and validation: We perform extensive unit testing of our code. Some of the data transformations rely on well-tested tools written in C++ such as Max Planck's Climate Data Operators and GDAL (depended upon by *rasterio*). For such cases, we rely on snapshot and regression testing to ensure the results are reproducible. Unit tests are run automatically on every code change using continuous integration with GitHub Actions. Regression testing is performed using cached data. Data transformations for a subset of climate variables (temperature and precipitation), as well as all socio-demographic variables are manually validated by experts in climate data. Range validation is performed (Table S3) and valid_min and valid_max attributes set on netCDF variables to indicate permissible ranges to downstream applications.

Extensibility: DART-Pipeline is written to be extensible and adaptable for future work that includes additional variables. Our code is open-source under the MIT license (https://opensource.org/license/mit), allowing others to build upon our work. The bias-correction component is available as a separate module (https://github.com/DART-Vietnam/dart-bias-correct) as it is GPL-3.0 licensed (https://opensource.org/license/gpl-3-0) due to a dependency. The forecast correction component is available as part of another repository (https://github.com/DART-Vietnam/dart-runner) that depends upon dart-bias-correct and the main DART-Pipeline. In addition, we have refactored out foundational components that are of use beyond the pipeline, such as manipulating rasters in memory, performing zonal statistics with a shapefile, and fetching reanalysis data from ECMWF into its own open-source package *geoglue* (https://geoglue.readthedocs.io) that users can use for tasks beyond this pipeline. Within the pipeline, code for each data source is in a single Python submodule, allowing extensibility to other climate, socio-demographic and epidemiological datasets in the future.

Accessibility of data sources: All primary data sources for DART Pipeline are open access, either through direct downloads or via a free-to-use API. Data sources required for bias correction are not open access – we provide links to instructions to get access. Users can run the pipeline without bias correction with entirely open-access data. For the purposes of reproducibility, we maintain a private AWS S3 bucket with source and intermediate files created during the processing of the pipeline for the real-time data. While these data cannot be shared, users can reproduce the output by re-downloading data and running the pipeline. ERA5 reanalysis data can be obtained by registering and logging in. To enable easier access, we prompt the user to create API tokens and allow easy entry of these tokens when the pipeline is run, along with documentation about licensing and redistribution.

Installation and compatibility: DART-Pipeline has been tested on macOS Apple Silicon, macOS x86_64 architecture and Linux x86_64. The minimum Python version supported is 3.11. Users also need to download the Climate Data Operators program (*cdo*) separately which is required by the resampling step[29]. A step-by-step guide on how to install and use the pipeline is provided in our documentation.

Use case: For the purpose of testing the pipeline, we selected a comprehensive list of meteorological variables from ERA5 and socio-demographic data from WorldPop and Meta and processed them for a dengue forecasting pipeline in Vietnam. Dengue epidemiology in Vietnam is highly heterogeneous with hyper-endemicity and year-round transmission in the southern part of the country and long-term emerging and high seasonality in the northern part[30].

We chose HCMC as the exemplar study region. HCMC is located in the southern part of the country. It is the economic capital and most populous city of Vietnam, with large seasonal dengue outbreaks during the rainy season. Like many large urban centres in endemic areas, outbreaks vary in size and peak timings from year to year depending on climate and socio-demographic factors[31]. Dengue control in the city relies largely on targeted vector control managed through the HCMC Center for Disease Control (HCDC); however, due to resource constraints, control efforts are carried out in response to local outbreaks in the city and not proactively. In such settings, forecasting at the spatial level for which vector control and hospital patients are managed, i.e. district level, can potentially improve the effectiveness of these vector control measures.

As the epidemiological data cannot be shared outside of Vietnam due to legal data sharing restrictions, we developed the pipeline to be locally deployable and open-source to facilitate adoption in the local context. The epidemiological data for HCMC are available from 2000 to 2022. The data, which contains residential addresses of admitted patients, is then aggregated to the district level. The meteorological and socio-economic data listed in Table S1 were aggregated to the same resolution using the tool described in this paper.

Data from 2020 to 2022 are excluded for subsequent analyses because (i) COVID-19 lockdowns in the city during 2020 and 2022 affected dengue virus transmission and reporting due to healthcare resources that were diverted elsewhere (Figure S2), and (ii) a new sub-city (Thu Duc city) was established at the end of 2020[32], rearranging the spatial boundaries of districts within HCMC. Data from 2000 are also excluded due to concerns of data reporting quality and to keep temporal ranges in-line with other data. In summary, we use aggregated district-level dengue incidence time series data from 2001 to 2019 (inclusive). We illustrate end-to-end operation of the pipeline by training a SARIMAX model on synthetic dengue incidence data (Figure S3).

Discussion

As climate-driven infectious diseases are increasing in their frequency and intensity, forecasting models play an important role in the design of pre-emptive and reactive interventions. For climate-sensitive diseases, there is currently a gap in the ability to rapidly integrate heterogeneous large-scale datasets into probabilistic models for timely and continuous risk assessment[10]. To address this gap, we developed a scalable, modular, and open-access pipeline for the ingestion, integration, correction and aggregation of tabular (wealth index)

and gridded imagery data. The tool is modular by design and thus extendable to other data types and geographical contexts and improves the ability to perform interdisciplinary analyses of data before, during and ahead of outbreaks of infectious diseases. Beyond that, however, the pipeline is applicable to other research applications that involve the use of weather, climate, and socio-economic data sources such as monitoring changes and drivers in biodiversity.

Best practices in data aggregation and integration for informing control and prevention strategies include several priority areas, such as interdisciplinary collaborations, protocols for sustainable data, reproducible and user-friendly digital tools. Similarly, Ryan *et al.*[10] recommended that tools for climate-sensitive infectious diseases should incorporate both climate and epidemiological data, be transparently described and validated, be named, and be accessible and open-access. We sought to incorporate such practices throughout our research, leading to our framework which we believe reflects the 3-U (useful, usable, and used) research framework proposed for adoptable and sustainable digital prediction tools[33]. In particular, our common data model for integration of data types facilitates downstream analyses of these data together with locally acquired epidemiological data, thus enhancing comparability across settings (as per the FAIR principles[34]). While we have made an effort to make as much data available via this pipeline, not all processing can be automated and data be redistributed. For example, a user must request API access to the ERA5 climate data to deploy the model locally. Further, restrictions on redistribution of data might exist and change in the future, even when their intended use is for research purposes only. We therefore cannot provide data with redistribution restrictions via this pipeline (see Table S1 for licenses).

The selection of variables in this first version of the pipeline is not exhaustive. However, other variables such as land cover type or mosquito suitability can be easily added to the pipeline and we provide instructions on how to do so in our documentation (https://dart-pipeline.readthedocs.io/en/latest/reference/custom_metrics.html). Future work may seek to incorporate automated or semi-automated (with user input) variable selection and thus, model selection procedures for dengue prediction models. These selection procedures would likely involve statistically rigorous metrics, e.g. predictive/information criteria and scoring rules[35–37], which could allow different models for different users, regions, and times of the year. Other future work may incorporate explainability metrics[38–40] to automatically capture how individual covariates (both climatic and socio-demographic) influence model predictions, thus providing further transparency for the user in understanding data-driven model processes (data ingestion, model training, and prediction) and important epidemic dynamics and drivers.

While we acknowledge that data integration is only one part of gaining a more comprehensive understanding into the complex transmission dynamics of climate-sensitive infectious diseases, they are critical for robust downstream analyses and reproducibility, especially in settings where automation and re-generation of prediction/forecasting results are required, such as probabilistic forecasting for an outbreak. However, epidemiological data are frequently not openly accessible and biases in them are often unique and depend on the local surveillance systems. Our tool enables standardised integration of data types for drivers of transmission for local partners but today does not accommodate tools for correcting for possible delays in reporting or missing data. However, possible extensions can be easily integrated in future releases[41,42]. We have developed our pipeline to be flexible and modular in order to facilitate integration in analysis pipelines. For that we plan to formally integrate the pipeline into applications of dengue modelling in Vietnam and other countries globally.

Multi-modal data integration is a necessary step for robust analyses of disease outbreaks. Our tool is a first step in a platform for data integration and analyses that we believe contributes to future epidemic and pandemic preparedness[43], including monitoring and mitigating the impact of climate on infectious diseases.

Ethics statement

This study did not require ethical approval as it did not involve human participants, human data, or animals. All data used in the pipeline were derived from climate and population datasets. Therefore, no ethical issues are associated with the use of the data.

Data and software availability

Source data

Source data for the pipeline is publicly available, see below

Software availability

Source code for the DART-Pipeline is available on GitHub (https://github.com/kraemer-lab/DART-Pipeline) under MIT license. Source code for dart-bias-correct (https://github.com/DART-Vietnam/dart-bias-correct) and dart-runner (https://github.com/DART-Vietnam/dart-runner) are available on GitHub under the GPL-3.0 license.

Data availability

Most source datasets are freely available, except those for forecast bias correction that relies on historical forecast data from ECMWF's MARS service. ERA5 data at https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=overview is downloaded using the *cdsapi* Python package (user authentication required). ECMWF weather forecast data is accessed from ECMWF open data website (https://data.ecmwf.int/forecasts/) using the *ecmwf.opendata* Python package. Worldpop population density data is available at https://hub.worldpop.org/geodata/listing?id=75. Meta relative wealth index is available at https://data.humdata.org/dataset/relative-wealth-index. Geospatial data is obtained from the GADM project at https://gadm.org. Code to reproduce figures is at https://doi.org/10.5281/zenodo.17433818. The data that support the findings of this study are available from authors upon reasonable request by emailing abhishek.dasgupta@dtc.ox.ac.uk or iago.perezfernandez@eng.ox.ac.uk.

Extended data

This article has supplementary information located at https://doi.org/10.5281/zenodo.17565722, containing the following figures and tables:

Figure S1. Map of administrative boundaries of the 24 districts (Global Administrative Level 2) of the HCMC province (Global Administrative Level 1), Vietnam, in 2020.

Figure S2. Reported dengue incidence in HCMC, Vietnam, from 2000 to 2022.

Figure S3. Predicted dengue incidence from SARIMAX models trained using synthetic case incidence data and ERA5 data zonally aggregated from 2001—2019.

Table S1. Resampling methods and aggregation schemes for each dataset and variable, used to achieve a common resolution for dengue modelling.

Table S2. Formulae for spatial aggregation used by exactextract

Table S3. Automated validation checks by variable

Acknowledgements

We thank all individuals who have given feedback at various stages of the project including members of the DART project (https://www.dartdengue.org/). Their feedback and insights were invaluable.

References

1. Tsui JLH, Pena RE, Moir M, *et al.*: **Impacts of climate change-related human migration on infectious diseases.** *Nat Clim Chang.* 2024; **14**: 793–802.
Publisher Full Text

2. Poongavanan J, Lourenço J, Tsui JLH, *et al.*: **Dengue virus importation risks in Africa: a modelling study.** *Lancet Planet Health.* 2024; **8**(12): e1043–e1054.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Kraemer MUG, Reiner RC Jr, Brady OJ, *et al.*: **Past and future spread of the arbovirus vectors *Aedes aegypti* and *Aedes albopictus*.** *Nat Microbiol.* 2019; **4**(5): 854–863.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Mora C, McKenzie T, Gaw IM, *et al.*: **Over half of known human pathogenic diseases can be aggravated by climate change.** *Nat Clim Chang.* 2022; **12**(9): 869–875.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Chen Z, Tsui JLH, Gutierrez B, *et al.*: **COVID-19 pandemic interventions reshaped the global dispersal of seasonal influenza viruses.** *Science.* 2024; **386**(9): eadq3003.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Wesolowski A, Qureshi T, Boni MF, *et al.*: **Impact of human mobility on the emergence of dengue epidemics in Pakistan.** *Proc Natl Acad Sci U S A.* 2015; **112**(38): 11887–11892.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Harish V, Colón-González FJ, Moreira FRR, *et al.*: **Human movement and environmental barriers shape the emergence of dengue.** *Nat Commun.* 2024; **15**(1): 4205.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Gibb R, Colón-González FJ, Lan PT, *et al.*: **Interactions between climate change, urban infrastructure and mobility are driving dengue emergence in Vietnam.** *Nat Commun.* 2023; **14**(1): 8179.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Lowe R, Codeço CT: **Harmonizing multisource data to inform Vector-Borne Disease risk management strategies.** *Annu Rev Entomol.* 2025; **70**(1): 337–358.
PubMed Abstract | Publisher Full Text

10. Ryan SJ, Lippi CA, Caplan T, *et al.*: **The current landscape of software tools for the climate-sensitive infectious disease modelling community.** *Lancet Planet Health.* 2023; **7**(6): e527–e536.
PubMed Abstract | Publisher Full Text

11. Mills C, Donnelly CA: **Climate-based modelling and forecasting of dengue in three endemic departments of Peru.** *PLoS Negl Trop Dis.* 2024; **18**(12): e0012596.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Colón-González FJ, Soares Bastos L, Hofmann B, *et al.*: **Probabilistic seasonal dengue forecasting in Vietnam: a modelling study using superensembles.** *PLoS Med.* 2021; **18**(3): e1003542.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Mordecai EA, Caldwell JM, Grossman MK, *et al.*: **Thermal biology of mosquito-borne disease.** *Ecol Lett.* 2019; **22**(10): 1690–1708.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Mordecai EA, Paaijmans KP, Johnson LR, *et al.*: **Optimal temperature for malaria transmission is dramatically lower than previously predicted.** *Ecol Lett.* 2013; **16**(1): 22–30.
PubMed Abstract | Publisher Full Text

15. Tesla B, Demakovsky LR, Mordecai EA, *et al.*: **Temperature drives Zika virus transmission: evidence from empirical and mathematical models.** *Proc Biol Sci.* 2018; **285**(1884): 20180795.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Lowe R, Lee SA, O'Reilly KM, *et al.*: **Combined effects of hydrometeorological hazards and urbanisation on dengue risk in Brazil: a spatiotemporal modelling study.** *Lancet Planet Health.* 2021; **5**(4): e209–e219.
PubMed Abstract | Publisher Full Text

17. Kraemer MUG, Perkins TA, Cummings DAT, *et al.*: **Big city, small world: density, contact rates, and transmission of dengue across Pakistan.** *J R Soc Interface.* 2015; **12**(111): 20150468.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Wimberly MC, Davis JK, Hildreth MB, *et al.*: **Integrated forecasts based on public health surveillance and meteorological data predict West Nile virus in a high-risk region of North America.** *Environ Health Perspect.* 2022; **130**(8): 87006.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Betts AK, Chan DZ, Desjardins RL: **Near-surface biases in ERA5 over the Canadian prairies.** *Front Environ Sci.* 2019; **7**: 478102.
Publisher Full Text

20. Hersbach H, Bell B, Berrisford P, *et al.*: **The ERA5 global reanalysis.** *Q J R Meteorol Soc.* 2020; **146**(730): 1999–2049.
Publisher Full Text

21. Cannon AJ, Sobie SR, Murdock TQ: **Bias correction of GCM precipitation by quantile mapping: how well do methods preserve changes in quantiles and extremes?** *J Clim.* 2015; **28**(17): 6938–6959.
Publisher Full Text

22. Main L, Sparrow S, Weisheimer A, *et al.*: **Skilful probabilistic medium-range precipitation and temperature forecasts over Vietnam for the development of a future dengue early warning system.** *Meteorol Appl.* 2024; **31**(4): e2222.
Publisher Full Text

23. REMOCLIC: **VnGP - Vietnam Gridded Precipitation dataset (0.25°× 0.25°), Data Integration and Analysis System (DIAS).** 2016.

24. Perez I, Sparrow S, Weisheimer A, *et al.*: **Relative humidity verification over Vietnam for a Dengue warning system development.** *Under review in Meteorological Applications.*

25. Ith S, Seposo X, Phy V, *et al.*: **Extreme weather events and dengue in Southeast Asia: a regionally-representative analysis of 291 locations from 1998 to 2021.** *PLoS Negl Trop Dis.* 2025; **19**(9): e0012649.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Chen Y, Xu Y, Wang L, *et al.*: **Indian Ocean temperature anomalies predict long-term global dengue trends.** *Science.* 2024; **384**(6696): 639–46.
PubMed Abstract | Publisher Full Text

27. **GADM maps and data.** *GADM*.
Reference Source

28. **Unidata**.
Reference Source

29. Schulzweida U: **CDO user guide.** *Zenodo.* [Preprint], 2023.
Publisher Full Text

30. Shepard DS, Undurraga EA, Halasa YA, *et al.*: **The global economic burden of dengue: a systematic analysis.** *Lancet Infect Dis.* 2016; **16**(8): 935–941.
PubMed Abstract | Publisher Full Text

31. Huber JH, Childs ML, Caldwell JM, *et al.*: **Seasonal temperature variation influences climate suitability for dengue, chikungunya, and Zika transmission.** *PLoS Negl Trop Dis.* 2018; **12**(5): e0006451.
PubMed Abstract | Publisher Full Text | Free Full Text

32. **Resolution on arrangement of commune-level and district-level administrative divisions and establishment of Thu Duc city affiliated to Ho Chi Minh city.** 2020.
Reference Source

33. Phung D, Colón-González FJ, Weinberger DM, *et al.*: **Advancing adoptability and sustainability of digital prediction tools for climate-sensitive infectious disease prevention and control.** *Nat Commun.* 2025; **16**(1): 1644.
PubMed Abstract | Publisher Full Text | Free Full Text

34. **FAIR Principles.** *GO FAIR.* 2017.
Reference Source

35. Mills C, Falconi-Agapito F, Carrera JP, *et al.*: **Multi-model approach to understand and predict past and future dengue epidemic dynamics.** *medRxiv.* 2024.
Publisher Full Text

36. Lopez VK, Cramer EY, Pagano R, *et al.*: **Challenges of COVID-19 case forecasting in the US, 2020–2021.** *PLoS Comput Biol.* 2024; **20**(5): e1011200.
PubMed Abstract | Publisher Full Text | Free Full Text

37. Mellor J, Tang ML, Jones O, *et al.*: **Forecasting COVID-19, influenza, and RSV hospitalizations over winter 2023–4 in England.** *Int J Epidemiol.* 2025; **54**(3): dyaf066.
PubMed Abstract | Publisher Full Text

38. Preece A: **Asking "Why" in AI: explainability of intelligent systems – perspectives and challenges.** *Intell Syst Account Finance Manag.* 2018; **25**(2): 63–72.
Publisher Full Text

39. Mills C, de Sousa G dos S, Lima Neto AS, *et al.*: **The time- and space-varying roles of human mobility in shaping urban dengue epidemics.** *medRxiv.* 2025.
Publisher Full Text

40. Coupland H, Scheidwasser N, Katsiferis A, *et al.*: **Exploring the potential and limitations of deep learning and explainable AI for longitudinal life course analysis.** *Res Sq.* 2025.
Publisher Full Text

41. Xiao Y, Soares G, Bastos L, *et al.*: **Dengue nowcasting in Brazil by combining official surveillance data and Google Trends information.** *PLoS Negl Trop Dis.* 2025; **19**(8): e0012501.
PubMed Abstract | Publisher Full Text | Free Full Text

42. Charniga K, Park SW, Akhmetzhanov AR, *et al.*: **Best practices for estimating and reporting epidemiological delay distributions of infectious diseases.** *PLoS Comput Biol.* 2024; **20**(10): e1012520.
PubMed Abstract | Publisher Full Text | Free Full Text

43. Kraemer MUG, Tsui JLH, Chang SY, *et al.*: **Artificial Intelligence for modelling infectious disease epidemics.** *Nature.* 2025; **638**(8051): 623–635.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 25 November 2025

https://doi.org/10.21956/wellcomeopenres.27767.r139748

✔ **Oliver Brady**
London School of Hygiene & Tropical Medicine, London, UK

Approved.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 22 November 2025

https://doi.org/10.21956/wellcomeopenres.27767.r139747

✔ **Kayode Oshinubi**
Northern Arizona University, Flagstaff, Arizona, USA

No furtehr comments for the authors.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: computational epidemiology, vector-borne diseases, infectious disease modeling

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 08 October 2025

https://doi.org/10.21956/wellcomeopenres.27294.r134098

Kayode Oshinubi
Northern Arizona University, Flagstaff, Arizona, USA

The authors have done a good job by developing this open access pipeline to help improve the predictability and forecast of vector-borne diseases, which are climate sensitive. As someone who is working on vector-borne disease and interested in using climate forecasts in my model, I find reading this article interesting and novel.
I do not have major revisions for the authors. I want the authors to clarify the following:
1. Most modelers like to use R software packages. Is there any plans to have an R version of this pipeline?
2. I strongly believe that there should be more details about the spatial resolution in which the datasets are available. The temporal resolution is very clear. For instance, in the USA, we have CBGs, counties, ZCTAs, etc. Do we have the data in this spatial scale? Furthermore, in terms of aggregation, how can we aggregate to zip codes or ZCTA resolution?

Aside from the above-stated minor revisions, I am strongly in support of this article, which will be an added value to the research community, most especially those interested in using these datasets in their research.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: computational epidemiology, vector-borne diseases, infectious disease modeling

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 24 Oct 2025

Abhishek Dasgupta

The authors have done a good job by developing this open access pipeline to help improve the predictability and forecast of vector-borne diseases, which are climate sensitive. As someone who is working on vector-borne disease and interested in using climate forecasts in my model, I find reading this article interesting and novel.

Response: We thank the reviewer for their positive assessment of our work.

I do not have major revisions for the authors. I want the authors to clarify the following:

1. Most modelers like to use R software packages. Is there any plans to have an R version of this pipeline?

Response: We chose to use Python for its extensive library support for data science and machine learning and familiarity with the development team. While the library is written in Python, we provide a command line interface as the primary method of interaction with the pipeline that can be used to fetch and process data without knowledge of Python. In the dart-runner repository (https://github.com/DART-Vietnam/dart-runner), we provide an example integration with a machine learning model written in R.

2. I strongly believe that there should be more details about the spatial resolution in which the datasets are available. The temporal resolution is very clear. For instance, in the USA, we have CBGs, counties, ZCTAs, etc. Do we have the data in this spatial scale? Furthermore, in terms of aggregation, how can we aggregate to zip codes or ZCTA resolution?

Response: The primary spatial resolution in our study is determined by the Worldpop rasters used for population weighted aggregation that are 1km x 1km (see Methods, Spatial aggregation). Raster data is upsampled to this resolution and aggregated over GADM shapefile polygons at admin level 2. Users can provide their own shapefiles, so it would be possible to aggregate to the above mentioned resolutions provided appropriate shapefiles. Our pipeline is modular allowing extensibility to other data sources in the future.

Aside from the above-stated minor revisions, I am strongly in support of this article, which will be an added value to the research community, most especially those interested in using these datasets in their research.

Response: We thank the reviewer for their positive feedback and support..

Competing Interests: No competing interests were disclosed.

Reviewer Report 27 September 2025

https://doi.org/10.21956/wellcomeopenres.27294.r132631

Oliver Brady
London School of Hygiene & Tropical Medicine, London, UK

This article describes the development of a pipeline to ingest, bias correct and aggregate climate data- all necessary but fairly limited steps in the longer process of analysing, modelling and forecasting dengue case count data.

Major comments:
The authors need to make a clearer case for why existing GIS platforms do not met the needs of users and what advantages this software pipeline brings. Practically this could mean expanding the introduction to clearly articulate the gap that this software tries to fill and providing additional results from the user testing giving quantitative evidence of (e.g. faster run times, improved usability, etc, etc) ideally benchmarked against existing software.

The unit testing section is currently vaguely described with some important details missing. Specifically who were the users for the unit testing and how were they chosen, what data were collected and how was it analysed (general methods are mentioned, but specific details around what e.g. “regression testing” involves should be detailed).

No detail is provided about the relative wealth index from meta nor any necessary processing or validation steps. Typically, local users might want to use local census data to quantify socio-economic conditions. Is this possible within your framework.

While the software does have the capability to ingest epidemiological data its functionality is quite limited for any operational use of this type of data. No mention is included of how the system would handle data gaps, delays or anomalies- all common features of real-time epidemiological data that are necessary to address prior to developing an operational forecast. Is the intention for this software to only be used for historical analyses or are there plans to develop such functionality- and thus update the software description in this article?

The introduction pitches DART as a fairly broad one-stop-shop for “integration of diverse spatio-temporal epidemiological, socio-economic, and climatic datasets”, however this article only demonstrates processing of spatiotemporally complete raster reanalysis data- probably the least complex data type (in terms of completeness, resolution, timeliness, etc) that users might want to

work with many of the functionality needed to work with the other datasets missing. I wonder if presenting this as a more modest step in a wider modular framework that works towards this goal may be more appropriate. This might necessitate changing the title and parts of the introduction to focus more on the specific advantages of this pipeline for efficiently processing climate data.

Minor comments:
I'm fairly sure this functionality is included, but would be worth mentioning that users can provide their own shapefiles rather than relying on GADM and ISO3 standards. Administrative boundaries frequently change (as pointed out here in this paper) and are disputed so better for local users to have the flexibility to use their own shapefiles

Suggest making it clear in figure 1 that the dengue prediction model is not (currently?) part of this software

Could you clarify if the ERA5 and ECMWF weather forecasts are bias corrected separately or as a combined product? – I image the bias correction for long-term climate and short term weather forecasts could give quite different results.

While I appreciate the sensitivities that prevent sharing of any dengue data it might be useful to include a dummy dataset or one based on publicly available data to demonstrate end-to-end functionality.

Is the rationale for developing the new software tool clearly explained?
Partly

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: Co-authored papers with author MUGK in past three years

Reviewer Expertise: Dengue epidemiology, modelling, spatial analysis.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 24 Oct 2025

Abhishek Dasgupta

The authors need to make a clearer case for why existing GIS platforms do not met the needs of users and what advantages this software pipeline brings. Practically this could mean expanding the introduction to clearly articulate the gap that this software tries to fill and providing additional results from the user testing giving quantitative evidence of (e.g. faster run times, improved usability, etc, etc) ideally benchmarked against existing software.

Response: We thank the reviewer for their feedback. In the introduction we have listed other GIS tools and reference a recent review that has provided an initial list (https://www.thelancet.com/journals/lanplh/article/PIIS2542-5196(23)00056-6/fulltext) with additional discussion about the importance of data integration here (https://www.annualreviews.org/content/journals/10.1146/annurev-ento-040124-015101). While some GIS platforms, such as Google Earth Engine exist, they focus on one particular data type, satellite imagery. We build on previous tools and stick together multiple processes including: Ingest: fetch raster data or read tabular (CSV) data, Preprocess: mask, compute derived variables; Aggregate: temporal rollups (weekly/monthly), group-by summaries, Join: Resample, zonal statistics; Output: push to database.

The unit testing section is currently vaguely described with some important details missing. Specifically who were the users for the unit testing and how were they chosen, what data were collected and how was it analysed (general methods are mentioned, but specific details around what e.g. “regression testing” involves should be detailed).

Response: We did not perform user testing (other than members of the team replicating the workflow in their own compute environments and performing it across multiple commonly used systems), unit testing here refers to testing units of the code. “Regression testing” refers to the process of ensuring that outputs of the pipeline do not change. These are best practices for developing research software (https://doi.org/10.1016/j.patter.2021.100206).

No detail is provided about the relative wealth index from meta nor any necessary processing or validation steps. Typically, local users might want to use local census data to quantify socio-economic conditions. Is this possible within your framework.

Response: In our pipeline, the Relative Wealth Index (RWI) is calculated by aggregating Meta’s wealth estimates, provided as tabular point-level data indexed by quadkeys, to administrative units using population-weighted averages. Population density at ~2.4 km resolution (matching the RWI data resolution) is taken from Meta’s population dataset and used as weights. Any missing or unmatched pixels are excluded through the inner join

during merging, so only overlapping data contribute to the aggregation. Within our framework, users can replace Meta's population layer with local census data aggregated to the required resolution, and this could be further expanded to directly incorporate tabular census data by linking it to administrative boundaries within the pipeline. We updated the manuscript to describe this process clearly and reference the relevant code module.

While the software does have the capability to ingest epidemiological data its functionality is quite limited for any operational use of this type of data. No mention is included of how the system would handle data gaps, delays or anomalies- all common features of real-time epidemiological data that are necessary to address prior to developing an operational forecast. Is the intention for this software to only be used for historical analyses or are there plans to develop such functionality- and thus update the software description in this article?

Response: The key contribution of this paper is in providing a standardised pipeline framework for pre-processing various kinds of data (tabular and gridded) that can be used in the analysis of climate-sensitive infectious diseases together with epidemiological data. We recognise that epidemiological data are a key part of any analysis but are often heterogeneous and inaccessible to researchers outside the agency or country where they are collected. Further, epidemiological data are biased in many unpredictable ways that would require deep understanding of the surveillance systems. However, some progress has been made to deal with reporting delays and gaps. Data imputation could be performed using time series foundation models (https://arxiv.org/abs/2411.07207) and delays in reporting has been discussed here (https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0012501 and https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1012520). Our framework is able to accommodate these tools and we have now mentioned possible extensions in the discussion: "Epidemiological data is often inaccessible and biases in them are often unique and depend on the local surveillance systems. Our tool enables standardised integration of data types for drivers of transmission for local partners but today does not accommodate tools for correcting for possible delays in reporting or missing data. However, possible extensions can be easily integrated in future releases."

The introduction pitches DART as a fairly broad one-stop-shop for "integration of diverse spatio-temporal epidemiological, socio-economic, and climatic datasets", however this article only demonstrates processing of spatiotemporally complete raster reanalysis data- probably the least complex data type (in terms of completeness, resolution, timeliness, etc) that users might want to work with many of the functionality needed to work with the other datasets missing. I wonder if presenting this as a more modest step in a wider modular framework that works towards this goal may be more appropriate. This might necessitate changing the title and parts of the introduction to focus more on the specific advantages of this pipeline for efficiently processing climate data.

Response: We thank the reviewer for their comment. Our tool is not only able to integrate raster re-analysis data but also tabular data (see RWI described above). See previous comment about integration of tools for epidemiological data. We have adapted the introduction and discussion to reflect that our tool is a step into the direction of a modular framework and thank the reviewer for this suggestion.

Minor comments: I'm fairly sure this functionality is included, but would be worth mentioning that users can provide their own shapefiles rather than relying on GADM and ISO3 standards. Administrative boundaries frequently change (as pointed out here in this paper) and are disputed so better for local users to have the flexibility to use their own shapefiles

Response: We have added text to the manuscript and online documentation (https://dart-pipeline.readthedocs.io/en/latest/workflow/using_custom_shapefiles.html) clarifying that users can provide custom configuration to use their own shapefiles.

Suggest making it clear in figure 1 that the dengue prediction model is not (currently?) part of this software

Response: We have updated figure 1 to indicate that the model is an external module.

Could you clarify if the ERA5 and ECMWF weather forecasts are bias corrected separately or as a combined product? – I image the bias correction for long-term climate and short term weather forecasts could give quite different results.

Response: ERA5 data and ECMWF weather forecasts are corrected using the same technique, but applied differently: First, ERA5 is corrected following the same approach as in Refs 22,24 of the main article, this is, the straightforward application of quantile mapping using Vietnam gridded precipitation data as reference. Second, real time weather forecasts use ERA5 as reference and in contrast to the latter, they have an extra dimension compared to ERA5 data as it shows different possible states of the atmosphere in the future (i.e., ensemble members). Each ensemble member was calibrated individually against ERA5. A sentence describing the different application of the bias-correction technique between weather forecast and ERA5 data was added to the main manuscript.

While I appreciate the sensitivities that prevent sharing of any dengue data it might be useful to include a dummy dataset or one based on publicly available data to demonstrate end-to-end functionality.

Response: We have added a dummy epidemiological dataset that utilises output data from the pipeline and a simple model (SARIMAX) to show end-to-end functionality at https://github.com/DART-Vietnam/DART-manuscript-figures. We have also added a Figure showing predictions from this simple model in the Supplementary Material.

Competing Interests: No competing interests were disclosed.