Part 2 Questions & Answers Session

Please type your questions in the Question Box. We will try our best to answer all your questions. If we don't, feel free to email Assaf Anyamba (assaf.anyamba@nasa.gov) or Michael Wimberly (mcwimberly@ou.edu).

Question 1: What are some of the limitations of remote sensing when applied to tracking climate-sensitive infectious diseases?

Answer 1: For climate-sensitive diseases, one of the challenges is that satellite observations do not always capture the most relevant environmental measurements. For example, we usually want measurements of near-surface air temperature to model vector-borne disease transmission, but satellite sensors measure land surface temperature (LST). Although air temperature and LST are often correlated, they are fundamentally different measurements (see question 3). Similarly, satellite measurements of precipitation are estimates that differ from the actual precipitation measured on the ground.

Another challenge is there are many climate sensitive diseases, but climate is not the only risk factor influencing their transmission cycles. Other information related to human exposure and vulnerability is important, as well as movements of human and non-human hosts that transport pathogens. Satellite data can help to measure some of these factors in addition to climate (see Dr. Loboda's presentation for some good examples), but other factors, such as human movement, are much harder to measure.

Despite these various challenges, it is interesting to note that in many cases, satellite observations are often as effective or more effective at measuring environmental risk factors for climate-sensitive diseases compared to in-situ measurements such as meteorological data from weather stations. See, for example, https://doi.org/10.1016/j.rse.2012.07.018 and https://doi.org/10.1093/jme/tjac145.

Question 2: What could be a better aggregation unit than Woreda? Sometimes aggregation to an Admin Unit does not take environmental variables, and other units give better results. What has been tested on these models? And which gives better results?

Answer 2: I would say that in the case of EPIDEMIA, woredas (districts) are actually a good resolution for looking at relationships with climate. The advantage of larger units like woredas is that with populations of 10's to 100's of thousands, they provide

relatively stable estimates of variability in disease incidence over weeks and years. This variation can then be correlated with climate fluctuations, which tend to be synchronized over broad areas.

Another option is to look at smaller spatial resolutions, like the kebele (village) level. This provides the opportunity to explore localized environmental factors that can affect the sensitivity of particular villages to climate fluctuations. The main limitation is that we don't have long-term data at this resolution over the entire country. Other limitations include lack of population data at this fine scale and difficulties ascertaining whether patients seeking care at a health center in a kebele are actually residents who were infected in the same kebele, or if they have travelled from neighboring kebeles to seek care. For an example of a kebele-level analysis that addresses some of these challenges, you can read this article: https://doi.org/10.1186/s12942-021-00282-0.

Question 3: Is land surface temperature a suitable proxy for air temperature?

Answer 3: LST is almost always different from air temperature, because it is strongly affected by land cover properties. So LST is usually not a good proxy if you are trying to estimate the air temperature at a particular location. However, LST anomalies (differences between the current temperature and the long-term mean for a given day or year and location) and air temperature anomalies are often strongly correlated, especially when summarized over large areas. To put it in plain language, LST does not tell you the exact air temperature. However, if LST is warmer than usual, then the air temperature is probably also warmer than usual (and vice versa). Here is a paper with some information on LST/Air temperature relationships in Ethiopia: https://doi.org/10.3390/s20051316.

Question 4: Wow this is great. Can I do this with TB?

Answer 4: The short answer is yes – you could definitely give it a try. I'm not an expert on TB but I took a quick look at the literature and there is evidence that climate factors such as temperature and atmospheric humidity can influence the survival and transmission of tuberculosis bacteria. However, as discussed in the answer to Question 1, keep in mind that there are also non-climatic factors that influence the transmission of TB, so one should not expect climate factors to explain all (or even most) of the spatial and temporal patterns of TB cases.

Question 5: Are the methods for this [time series model] published?

Answer 5: Yes, I can suggest a few relevant publications.

Davis, J. K., Gebrehiwot, T., Worku. M., Awoke, W., Mihretie, A., Nekorchuk, D., and M. C. Wimberly. 2019. A genetic algorithm for identifying spatially-varying environmental drivers in a malaria time series model. Environmental Modelling and Software 119: 275-284. https://doi.org/10.1016/j.envsoft.2019.06.010

Midekisa A., B. Beyene, A. Mihretie, E. Bayabil, M. C. Wimberly. 2015. Seasonal associations of climatic drivers and malaria in the highlands of Ethiopia. Parasites & Vectors 8: 339. https://doi.org/10.1186/s13071-015-0954-7

Midekisa, A., G. Senay, G. M. Henebry, P. Semuniguse, and M. C. Wimberly. 2012. Remote sensing-based time series models for malaria early warning in the highlands of Ethiopia. Malaria Journal 11: 165. https://doi.org/10.1186/1475-2875-11-165

Question 6: Does the distributed lag effect link the daily value of precipitation from 180 days prior to a daily malaria measurement? Or is there some aggregation of precipitation effects over longer time periods (weekly, monthly)?
Answer 6: You can think about a distributed lag as a weighted effect of the values of precipitation (or other weather variable) over the prior 180 days. Not all the days are having an equal effect. The response variable might be sensitive to recent precipitation in the past few days, precipitation that happened months in the past, or the cumulative effects of precipitation averaged over several weeks or months. These effects can be conceptualized as a set of weights for each of the previous 180 days, which are determined in the model fitting process, and can give more emphasis on more recent or older values depending on the underlying relationships in the data. With the distributed lag approach the analyst does not have to guess the lag structure beforehand – the algorithm determines the lag structure that provides the best fit between the climate predictor and the response.

Question 7: Do we need any specific software to complete the homework?
Answer 7: No.

Question 8: How much does it cost to use EPIDEMIA?
Answer 8: EPIDEMIA is free. The main cost is the time that must be invested. Some time is required to learn how to use the tools, and a considerable amount of time is often required to acquire, process, and format the necessary data to drive the models.

You can find all the links to relevant GitHub packages and other resources at the following website:
https://ecograph.net/epidemia/

Question 9: Can EPIDEMIA be used with python instead of R?
Answer 9: The short answer is no, because the code is written in R. However, it might be possible to write a Python script that calls and runs EPIDEMIA in R. It might also be possible to rewrite the entire code case in Python (we are actually thinking about doing this for the next implementation), but the limiting factor would be whether we can use similar modeling packages in Python as we can in R.

Question 10: How is the GEE data extraction in your case study different from just downloading CDS data using a bounding box? Just already tabulated climate data.
Answer 10: I expect that you could obtain similar satellite data products through the Copernicus Data Store. We used Google Earth Engine to acquire data because we were able to develop a simple app that automated most of the steps for end users. More details are available in this paper: https://doi.org/10.1038/s41597-022-01337-y.

Question 11: Estimating lag is often not the only complexity for this type of time series forecasting with covariates. Understanding if what matters is the sum of rainfall (for instance) over the previous period, the max, the deviation from the mean, is usually complex. I didn't catch if this is part of what you analyzed? One approach is usually to use feature generation to extract several statistics of the covariates over variable periods pre-date of interest. Is this something you have looked into?
Answer 11: I think the answer here is yes and no. We do use anomalies (deviations from the long-term mean on any given day or week) of remotely-sensed climate variables in all of our models because we have found that these variables generally perform better for predicting outbreaks than simply using the raw satellite indices. We have not had much success in using extreme values (e.g., maximum precipitation) in our models, but this is something that we explored a while ago and it would be interesting to revisit with some newer analytical approaches.

Question 12: In the distributed lag model, how is the relative risk calculated in case of lags? Does the calculated risk of (i+1)th lag accounts for effects of ith lag? Is it cumulative?

Answer 12: See the answer to question 6. The effects are cumulative over the previous 180 days, with the relative effects at different lags determined in the model fitting process.

Question 13: I would like to ask for your guidance on the best approach to working with 23 year time series data. Specifically, I am analyzing leprosy cases in a province in northern Argentina in relation to environmental variables such as NDVI and LST. Given the volume of data, I find it difficult to determine the most appropriate analytical strategy. Any recommendations or guidance will be of great help.

Answer 13: I don't know enough about your problem or these data to give specific recommendations, but I can offer a few general suggestions. One general challenge with time series data is that there are multiple signals embedded, including long-term trends, interannual variations, seasonal cycles, and random noise. I think that time series decomposition methods are often very useful for visualizing the different signals in a dataset. Cross-correlation plots are also helpful for visualizing the lagged associations between climate predictors and the disease response. I strongly recommend doing lots of visualization and making sure you understand the basic structures and relationships in your dataset before diving into more complex models.

Question 14: Can the R code in EPIDEMIA be customized to other diseases?

Answer 14: In principle, yes. None of the underlying data processing or modeling approaches are specific to a single disease. However, it would take some work to redo the reports, which are specifically developed to visualize two types of malaria caused by *Plasmodium falciparum* and *Plasmodium vivax*.

Question 15: Who owns the data? Does Ethiopia share the weekly model forecasts publicly?

Answer 15: The Ethiopian government owns the data. Our approach has been to build tools to help the agencies to use "in house" with their own systems and data. In the past the reports have been used internally by regional and national public health agencies for decision making, but have not been shared publicly.

Question 16: Which specific spectral features or narrow-band indices derived from hyperspectral data are most effective for detecting subtle changes in vector habitat quality (e.g., changes in water quality, specific host plant health) that are driven by climate variability?

Answer 16: I do not have much experience with hyperspectral imagery, so I am not sure. But I think it would be less useful for detecting climate signals and more useful at higher spatial resolution to identify specific features such as larval habitats for vectors. Using hyperspectral data to eluciate host plant health could also be very useful for assessing vulnerability to crop pests.

Question 17: In the time series reports providing details on Woreda, the bottom chart shows three different parameters (temperature, humidity, and precipitation). When there is an increase in precipitation and decrease in land surface temperature, it co-relates to an increase in Malaria outbreak. Has another study been conducted to narrow down to what are the chances that a single parameter (e.g., Humidity/Precipitation/Temperature) played a more significant role in the increase of Malaria?
Answer 17: One thing that we have consistently found is that the relative importance of these predictors varies geographically. For example, temperature is a more important constraint in cooler high-elevation regions than warmer low-elevation regions, and precipitation is a more important constraint in drier regions than wetter regions. For more information, I suggest checking out the references listed in the answer to Question 5.

Question 18: How can I collect data on air quality on human health?
Answer 18: This is outside of my area of expertise, but I can recommend a review article that contains a lot of background information: https://doi.org/10.1146/annurev-biodatasci-110920-093120.
Many ARSET trainings are focused on air quality, which you can also review for more information. We recommend starting with "An Inside Look at how NASA Measures Air Pollution" for a basic overview.

Question 19: Shouldn't all remote-sensing climate data be calibrated before being used in a model?
Answer 19: Not necessarily. In many cases, it's not possible to calibrate remote sensing data because the measurement being collected from satellites don't always have a "gold standard" against which to calibrate. It's also worth noting that ground-based climate measurements from weather stations are necessarily the "correct" measurement, because the microclimates that influence vectors, hosts, and disease transmission are often very different from the conditions measured at these stations. As noted in the response to Question 1, satellite observations can actually be

better predictors of disease vectors and transmission than ground-based climate measurements because of their higher spatial precision and their ability to capture unique information about water bodies and other land surface characteristics.

Question 20: You mention that there were no other suitable weather data, but what about reanalysis datasets?
Answer 20: Yes, reanalysis datasets and interpolated weather datasets can also be used for this type of analysis. For example, we have used NLDAS data and GridMET data to develop predictive models of West Nile virus outbreaks in the United States, as described in these papers: https://doi.org/10.1289/EHP10287, https://doi.org/10.1016/j.actatropica.2018.04.028.

One limitation to reanalysis datasets is their coarse spatial resolution, which can range from 0.25-0.5 degrees for global products. This can be particularly problematic in mountainous environments where climate may vary over much finer scales because of orographic effects. Latency is another issue, as environmental data need to be available within a few days of the observations to be useful in an operational early warning system. Because of these issues, we decided to use higher-resolution, lower-latency satellite observations in EPIDEMIA.

Question 21: When we download climate data, it comes in grid (raster) format. However, health departments conduct actions in administrative areas (municipalities or counties). Is there a model or tool that can be used to aggregate/summarize climate data in these areas?
Answer 21: The simplest way to integrate these different types of data is through zonal statistics, a basic GIS operation in which the gridded data are summarized within each administrative polygon. This is part of the procedures that are carried out by the Google Earth Engine tool that we developed to obtain and process satellite data for EPIDEMIA: https://doi.org/10.1038/s41597-022-01337-y.

Question 22: How do you usually account for the implementation of public health interventions that prevent malaria in your models?
Answer 22: We can't do this directly, so we use long-term trends to indirectly account for the effects of interventions on malaria. The limiting factor is data availability. Information about interventions like long-lasting insecticide treated nets (LLITNs) and indoor residual spraying (IRS) is available in some locations, but not for all areas over the full temporal extent of our datasets. Furthermore, the data often provides

information on the associated supplies (number of LLITNs distributed or amount of insecticide sprayed), but not on the numbers of people actually using bednets or covered by IRS. This is an important topic for future research and model development that will require additional work to better monitor and measure these interventions.

Question 23: Is there any guideline on how to select minimum, maximum or mean of temperature and rainfall for malaria?

Answer 23: We used a model comparison approach based on Akaike's Information Criterion for variable selection. Our basic modeling structure included one temperature variable (min, max, or mean), one precipitation variable, and one spectral index (e.g., NDVI or NDMI). Various combinations of these variable types were compared to select the one that yielded the best fit. We also developed a generic algorithm that implemented this approach while simultaneously selecting clusters of woredas that resulted in the best model fit: https://doi.org/10.1016/j.envsoft.2019.06.010.