ERDC/GRL TN-20-1
February 2020

Spatial Downscaling Disease Risk Using Random Forests Machine Learning

By Sean P. Griffin

INTRODUCTION: Mosquito-borne illnesses are a significant public health concern, both to the Department of Defense (DoD) and the broader national and international public health community. A thorough grasp of the spatial distribution, patterns, and determinants of these diseases is needed to truly understand the threats they impose on public health (Pages et al. 2010). This information, when available, is often only at a sub-national to regional scale. Such data fails to meet tactical-level applications when diseases exhibit high local variation (Rytkonen 2004; Linard and Tatem 2012). Additionally, finer spatial resolution is also required to target disease burden successfully within the population and reduce exposure.

Previous research has applied spatial downscaling techniques to meet specific epidemiological study needs requiring more localized statistics. Examples include downscaling malaria incidence rates from regional to urban centers through multivariate regression, hand-foot-mouth disease from national to township levels using generalized linear models, and applying hierarchical Bayesian frameworks to develop 5 km gridded risk maps of malaria, *Plasmodium falciparum* (Gething 2012; Wang 2017; Altamiranda-Saavedra et al. 2018). While these studies were able to improve coarse-scale information, they still failed to meet a spatial resolution relevant to tactical-level epidemiological mapping applications or the processing time required to support time-sensitive operations.

This technical note (TN) describes a methodology aimed at improving coarse epidemiological information to much finer resolutions than achieved in previous studies by combining machine-learning with open-source, high-performance cloud computing. The result is a 1,000 meter (m) gridded raster product that provides a pixel-wise magnitude of risk that can be used directly for tactical mapping applications or serve as an input dataset for additional modeling applications.

DATA AND METHODS: The research presented in this TN focused on dengue, which is a mosquito-borne viral disease transmitted by female mosquitoes mainly of the species *Aedes aegypti*. This is the same vector responsible for transmitting chikungunya, yellow fever, and Zika infection. Dengue is endemic to the tropical belt and greatly influenced by rainfall, temperature, and unplanned rapid urbanization, with the severest form of disease being the leading cause of hospitalization and death among children and adults in Latin America and Asia (Brady et al. 2012). While oral prophylaxis can prevent mosquito-vector diseases such as malaria, there are no specific vaccines or antiviral treatments against dengue fever (Hesse et al. 2017). This lack of treatment not only puts local populations at risk, but also can adversely impact military operations.

Researchers at the Geospatial Research Laboratory (GRL) queried provincial-level dengue incidence rates at monthly intervals between 1998 and 2010 from Project Tycho, a global health research database maintained by the University of Pittsburgh (Panhuis et al. 2018). Cambodia served as the region of interest (ROI) due to the endemicity of dengue, high local variation in disease incidence, and availability of administrative-level statistics. The data were reformatted to Comma-separated values (CSV) and spatially joined in Esri ArcMap to the Large-Scale International Boundary (LSIB) shapefile (Humanitarian Information Unit 2017).

Google Earth Engine (GEE) served as the high-performance cloud computing (HPC) environment used to process monthly composites of environmental, demographic, and landscape covariates between 1998 and 2010. GEE combines a multi-petabyte catalog of satellite imagery and geospatial datasets with planetary-scale analysis capabilities that include vector and raster data processing, machine-learning classifiers, and time-series algorithms (Gorelick et al. 2017).

The methods in this research followed spatial downscaling principles found in similar studies that include improving coarse population and demographic data, and remotely sensed products such as precipitation, soil moisture, and surface temperature (Gaelle et al. 2016; Zhang et al. 2016; Ezzine et al. 2017; Pang et al. 2017). The downscaling methods use a statistical algorithm to determine a relationship between a coarser response variable and finer spatial resolution covariates. This study chose to apply the random forests (RF) regression algorithm because of its demonstrated ability to yield higher accuracy compared to linear modeling techniques, albeit more difficult to interpret than a traditional linear regression (Couronne et al. 2018). RF is an ensemble classifier that constructs multiple de-correlated random regression trees that are bootstrapped and aggregated using the mean predictions from all regression trees (Breiman 2001). RF models also provide a quantitative measurement of each variable's contribution to the regression output, which is useful in evaluating the importance of each variable concerning dengue prevalence and conditions that affect disease vector suitability.

In this case, the monthly dengue incidence rates previously compiled in ESRI ArcMap serve as the response variable. The monthly composites of environmental, landscape, and demographic geospatial data serve as the covariates used to develop a response function and model incidence rates to a user-defined output pixel size; this study selected 1000 meter output grid cells because it met the high-resolution criteria of previous fine-scale epidemiology studies (Sturrock et al. 2014; Delmelle et al. 2016). As previously stated, rainfall, temperature, and urbanization significantly affect the presence of dengue, primarily due to influences on habitat suitability for the mosquito vector, *Aedes aegypti*. The spatial covariates used in this study included precipitation, land surface temperature, normalized difference vegetation index (NDVI), population, land cover and land use, and elevation (Table 1, Figure 1).

Table 1. Spatial covariate types and data sources used in the epidemiological downscale model.

Type	Spatial Covariate	Source
Environmental	Precipitation	CHIRPS
	sum	
	mean	
	Land Surface Temperature (Day and Night)	MODIS
	min	
	mean	
	max	
	NDVI*	MODIS
Landscape	Elevation	SRTM
	Annual Land Cover Product	MODIS
Demography	Human Population	WorldPop

* Normalized Difference Vegetation Index, measure of vegetation cover and vigor

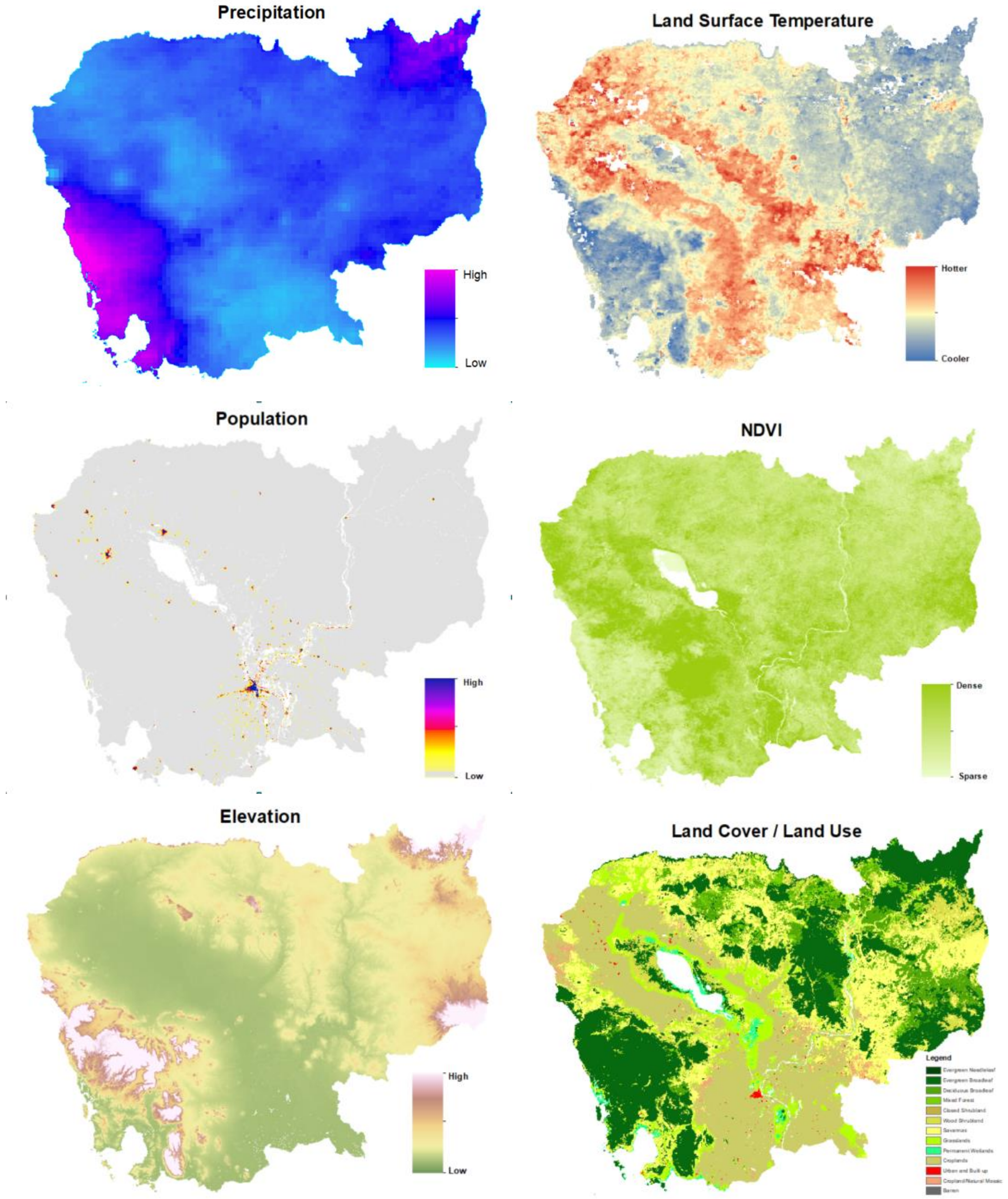


Figure 1. Spatial covariates representing environmental, landscape, and demographic determinants.

The spatial downscaling methodology is summarized in the following sequential steps:

1. Query and download administrative-level monthly dengue incidence rates from Project Tycho.
2. Spatially join dengue incidence rates to Large-scale International Boundaries (LSIB) shapefile and upload to Google Earth Engine (GEE) as a table asset.
3. Query environmental, landscape, and demographic spatial covariates in GEE and temporally reduce to monthly composites.

4. Select month and year to model.
5. Create a stratified sampling scheme in GEE and extract observed incidence rates (response variable) and environmental/landscape variables (covariates) for the date of interest.
6. Build random forest classifier using regression and run prediction.
7. Validate regression outputs by aggregating predicted grid cell values to the provincial boundary and compare to observed administrative-level incidence rates.

RESULTS

Spatial downscale output. Figure 2 provides a visual comparison between the gridded values derived from the RF regression downscale model and the observed provincial-level incidence rates for June 2010. The gridded output clearly shows a much higher spatial fidelity that meets any number of tactical and operational needs. The gridded output can serve as a disease risk map that could provide an understanding of the spatial variability in dengue and locations of higher risk to exposure. Also, the high-performance cloud-computing environment of GEE made it possible to develop a gridded model for the entire nation within minutes, a task that would be computationally intensive and time-consuming if duplicated in a desktop PC environment.

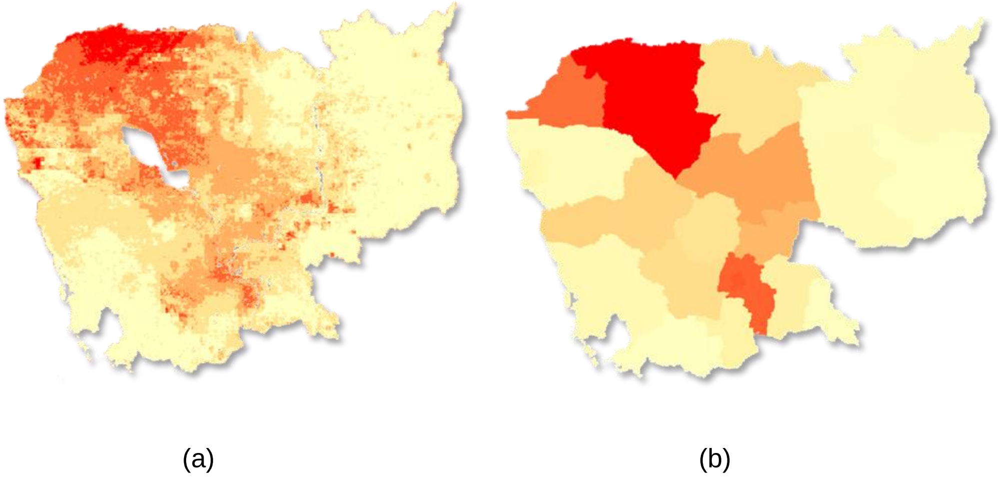


Figure 2. (a) Results of 1000-m downscaled product compared to (b) provincial-level statistics.

Figure 3 lists the RF spatial covariates in order of importance for June 2010. Population, temperature, vegetation cover, and precipitation were the most important variables, respectively, for describing the model, which coincides with epidemiological literature. The order of variable importance remained relatively consistent regardless of the chosen month and year.

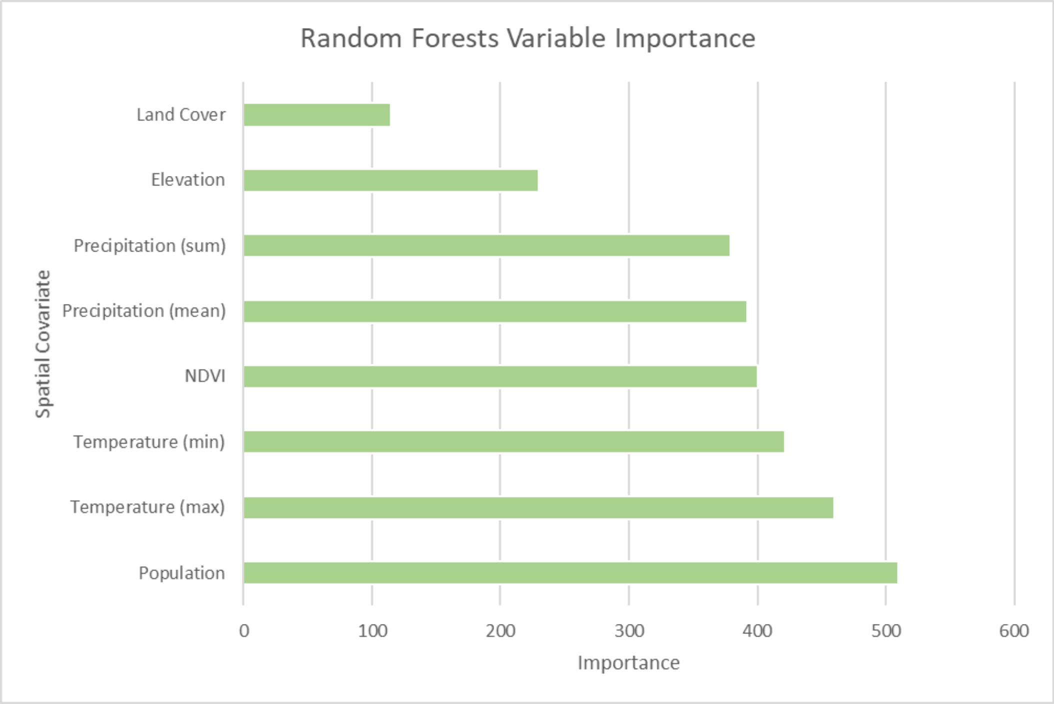


Figure 3. Spatial covariates in order of importance to the June 2010 downscale model.

Model validation. Figure 4 provides an example of model validation results for June 2010 using the spatial aggregation technique described in Step 8 of the methodology summary. Grid cell values of predicted disease incident rates were averaged within each administrative boundary and compared to the observed incidence rate for that given province. The absolute minimum and maximum difference between observed and downscaled data were 0.92 and 16.6 with the root mean square error (RMSE) being 5.64. A scatterplot was also used to compare observed and downscaled values yielding an r^2 of 0.87 (Figure 5).

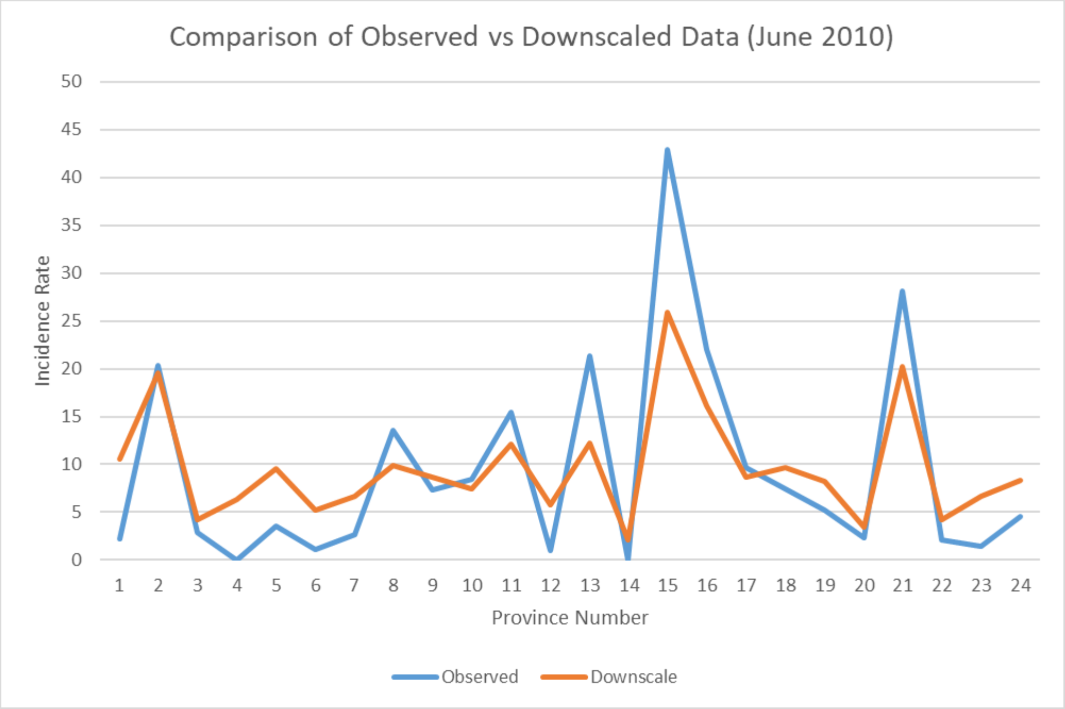


Figure 4. Comparison between observed and downscaled output per province.

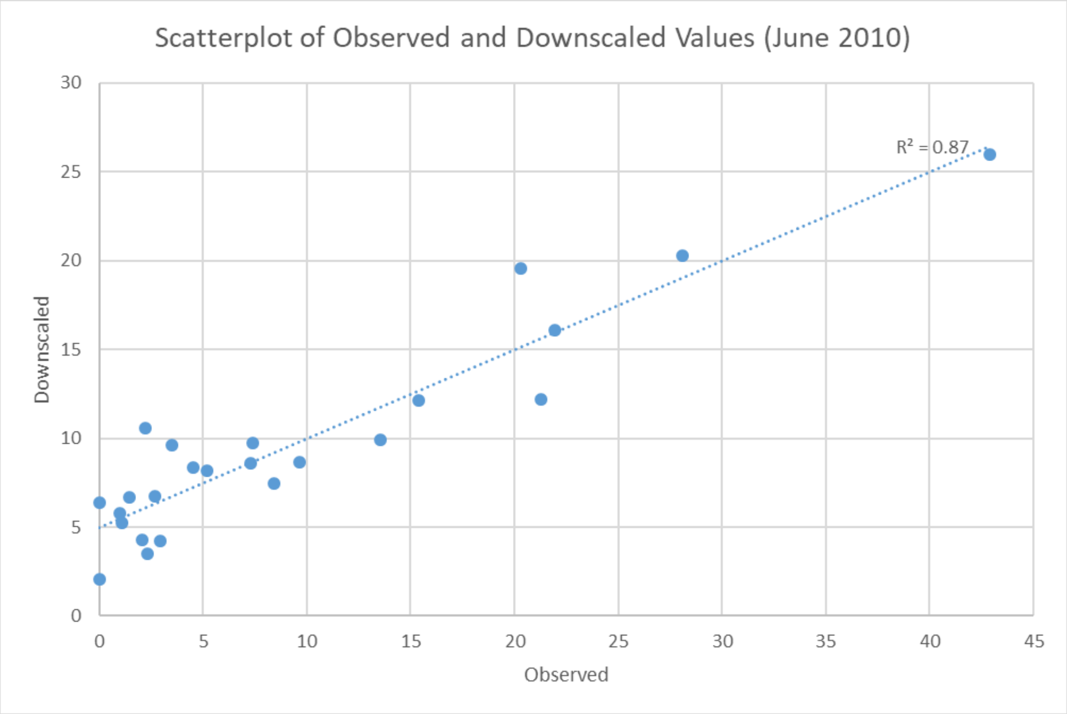


Figure 5. Scatterplot of observed and downscaled output per province.

SUMMARY AND CONCLUSIONS: Spatial downscale models were developed for each month between 1998 and 2010, totaling 156 geospatial disease risk products. All models showed significant agreement between downscaled and observed data with the highest RMSE being 10.25 and the lowest being 1.22. The lowest r^2 for the scatterplot comparisons was .72 and the highest was .94.

Random forests regression proved to be a high performing predictive algorithm that required little knowledge of machine-learning to achieve good results. However, the random forests model is not as easily interpretable as a traditional linear regression or classification/regression tree (CART), mainly due to the ensemble technique that creates hundreds of random, independent trees and combines the average into a single result.

GEE provided a high-performance computing environment that met the standards required for tactical and operational standards. Gridded products at 1,000 m spatial resolution were processed at national-levels within minutes as opposed to several hours on a desktop environment. Further advantages to GEE include reduced local resources, both related to computation and data accessibility. The main disadvantage to GEE is that it’s not aimed at the novice user since it requires programming knowledge of either JavaScript or Python.

Future contributing work to this study would explore the local spatio-temporal dynamics of the downscaled models. Dengue is known to be influenced by seasonal variables, such as precipitation and surface temperatures. Identifying strong temporal signals within a time-series could provide a further understanding of risk trends over time and possible associations with climate-disease teleconnections.

In conclusion, this study improved coarse, administrative-level disease data by downscaling to a 1000 m grid cell using random forests regression and spatial covariates. The generated output provides the level of tactical precision required to support Civil-Military Operations (CMO) targeting human health initiatives at a local scale. The output also provides a detailed geospatial product of disease risk that can be used to inform doctrine related to force health protection and force readiness during deployments.

REFERENCES

Altamiranda-Saavedra, M., X. Porcasi, C. M. Scavuzzo, and M. M. Correa. 2018. "Downscaling incidence risk mapping for a Colombian malaria endemic region." *Tropical Medicine and International Health* 23:10, 1101-1109. doi: 10.1111/tmi.13128

Brady, O.J., P. W. Gething, S. Bhatt, J. P. Messina, J. S. Brownstein, A. G. Hoen, C. L. Moyes, A. W. Farlow, T. W. Scott, and S. I. Hay. 2012. "Refining the global spatial limits of dengue virus transmission by evidence-based consensus." *PLoS Neglected Tropical Diseases* 6:8 e1760. doi: 10.1371/journal.pntd.0001760

Breiman, L. 2001. “Random Forests.” *Machine Learning* 45, pp. 5-32. doi: 10.1023/A:1010933404324

Couronné, R, P. Probst, and A. L. Boulesteix. 2018. “Random forest versus logistic regression: a large-scale benchmark experiment.” *Bioinformatics* 19:1. doi: 10.1186/s12859-018-2264-5

Delmelle, E. M., H. Zhu, W. Tang, and I. Casas. 2014. “A web-based geospatial toolkit for the monitoring of dengue fever.” *Applied Geography* 52, pp. 144-152. doi: 10.1016/j.apgeog.2014.05.007

Gething, P. W., I. R. F. Elyazar, C. L. Moyes, D. L. Smith, K. E. Battle, C. A. Guerra, A. P. Patil, A. J. Tatem, R. E. Howes, M. F. Myers, D. B. George, P. Horby, H. Wertheim, R. N. Price, I. Müeller, J. K. Baird, and S. I. Hay. 2012. "A long neglected world malaria map: Plasmodium vivax Endemicity in 2010." *PLoS Neglected Tropical Disease* 6:9. doi: 10.1371/journal.pntd.0001814

Hesse, E. M., L. J. Martinez, R. G. Jarman, A. G. Lyons, K. H. Eckels, R. A. De La Barrera, and S. J. Thomas. 2017. "Dengue virus exposures among deployed U.S. military personnel." *American Journal of Tropical Medicine and Hygiene* 96:5 pp. 1222–1226. doi:10.4269/ajtmh.16-0663

Hicham, E., B. Ahmed, O. Driss, and D. H. Moulay. 2017. "Downscaling of open coarse precipitation data through spatial and statistical analysis, integrating NDVI, NDWI, elevation, and distance from sea." *Advances in Meteorology* 17. doi: 10.1155/2017/8124962

Joint Publication 3-29. 2019. Foreign Humanitarian Assistance. Accessed August 1, 2019. https://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/jp3_29.pdf

Linard, C., and A. J. Tatem. 2012. "Large-scale spatial population databases in infectious disease research." *International Journal of Health Geographics* 11:7. doi: 10.1186/1476-072X-11-7

Nicolas, G., T. P. Robinson, G. R. William Wint, G. Conchedda, G. Cinardi, and M. Gilbert. 2016. "Using random forest to improve the downscaling of global livestock census data." *PLoS One* 11:3. doi: 10.1371/journal.pone.0150424

Pages, F., M. Faulde, E. Orlandi-Pradines, and P. Parola. 2010. " The past and present threat of vector-borne diseases in deployed troops." *Clinical Microbiology and Infectious Diseases* 16: 209–224. doi:10.1111/j.1469-0691.2009.03132.x

Pang, B., J. Yue, G. Zhao, and Z. Xu. 2017. "Statistical downscaling of temperature with the random forest model." *Advances in Meterology* 17. doi: 10.1155/2017/7265178

Rytkonen, M. J. 2004. "Not all maps are equal: GIS and spatial analysis in epidemiology." *International Journal of Circumpolar Health* 63:1, pp. 9-24. doi: 10.3402/ijch.v63i1.17542

Sturrock, H., J. M. Cohen, P. Keil, A. J. Tatem, A. L. Menach, N. E. Ntshalintshali, M. S. Hsiang, and R. D. Gosling. 2014. "Fine-scale malaria risk mapping from routine aggregated case data." *Malaria Journal* 13:421. doi: 10.1186/1475-2875-13-421

Wang, J. X., M. G. Hu, S. C. Yu, and G. X. Xiao. 2017. "Downscaling research of spatial distribution of incidence of hand foot and mouth disease based on area-to-area Poisson Kriging method." *Europe PMC* 38:9, pp. 1201-1205. doi: 10.3760/cma.j.issn.0254-6450.2017.09.012

Zhang, Q., P. Shi, V. P. Singh, K. Fan, and J. Huang. 2016. "Spatial downscaling of TRMM-based precipitation data using vegetative response in Xinjiang, China." *International Journal of Climatology* 37:10, pp. 3895-3909. doi: 10.1002/joc.4964

DATA SOURCE DOCUMENTATION

CHIRPS Pentad: Climate Hazards Group InfraRed Precipitation with Station Data (version 2.0 final) Funk, Chris, Pete Peterson, Martin Landsfeld, Diego Pedreros, James Verdin, Shraddhanand Shukla, Gregory Husak, James Rowland, Laura Harrison, Andrew Hoell & Joel Michaelsen. "The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes". Scientific Data 2, 150066. doi:10.1038/sdata.2015.66 2015.

LSIB: Large Scale International Boundary Polygons, Simplified. U.S. Department of State, Office of the Geographer at https://catalog.data.gov/dataset/global-lsib-lines-simplified-2017mar30

MCD12Q1.006 MODIS Land Cover Type Yearly Global 500m. Friedl, M., D. Sulla-Menashe. *MCD12Q1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006*. 2019, distributed by NASA EOSDIS Land Processes DAAC, https://doi.org/10.5067/MODIS/MCD12Q1.006.

MOD11A2.006 Terra Land Surface Temperature and Emissivity 8-Day Global 1km. Wan, Z., S. Hook, G. Hulley. MOD11A2 MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 1km SIN Grid V006. 2015, distributed by NASA EOSDIS Land Processes DAAC, https://doi.org/10.5067/MODIS/MOD11A2.006

MOD13Q1.006 Terra Vegetation Indices 16-Day Global 250m. Didan, K. MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006. 2015, distributed by NASA EOSDIS Land Processes DAAC, https://doi.org/10.5067/MODIS/MOD13Q1.006

SRTM: Digital Elevation Data 30m. Farr, T.G., Rosen, P.A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., and Alsdorf, D.E., 2007, The shuttle radar topography mission: Reviews of Geophysics, v. 45, no. 2, RG2004, at https://doi.org/10.1029/2005RG000183. https://developers.google.com/earth-engine/datasets/catalog/USGS_SRTMGL1_003.

Van Panhuis, W., A. Cross, D. Burke, and M. Choisy. 2018. Counts of Dengue reported in CAMBODIA: 1980-2011 (version 2.0, April 1, 2018): Project Tycho data release, DOI: 10.25337/T7/ptycho.v2.0/KH.38362002

SOFTWARE DOCUMENTATION

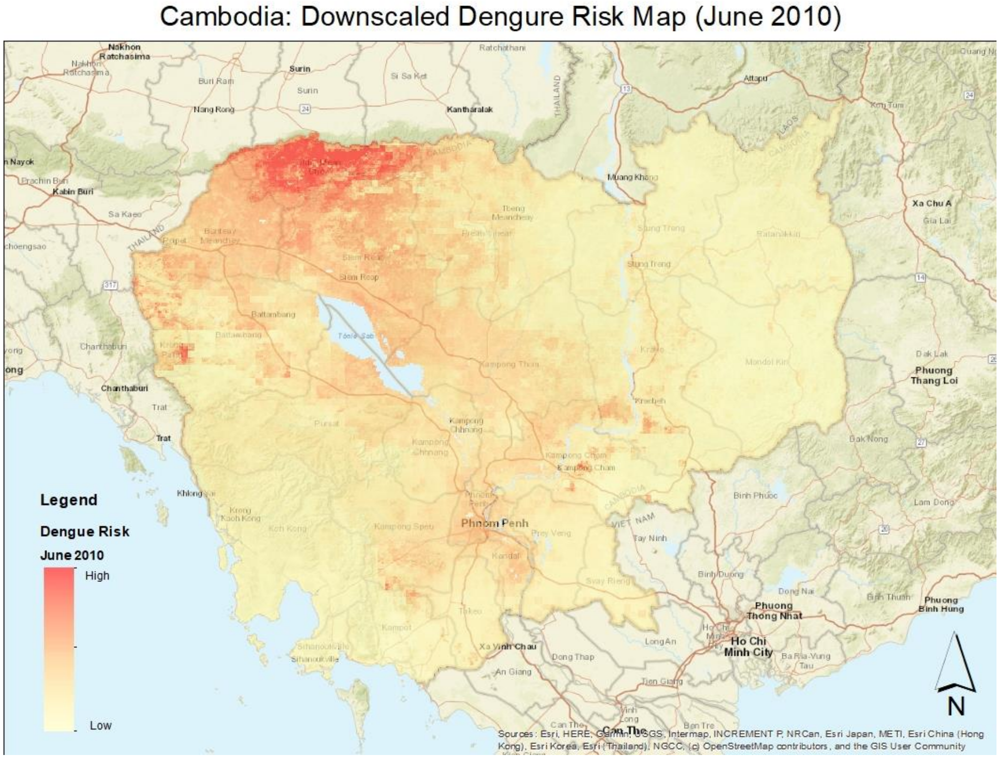
ArcGIS® software by Esri. ArcGIS® and ArcMap™ are the intellectual property of Esri and are used herein under license. Copyright © Esri. All rights reserved. For more information about Esri® software, please visit www.esri.com."

ENVI: Environment for Visualizing Images. L3Harris Geospatial Solutions. https://www.harrisgeospatial.com/Software-Technology/ENVI

Gorelick, N., M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment.

APPENDIX: Distribution Map

Spatially downscaled dengue risk map for Cambodia, June 2010.



Cambodia: Downscaled Dengure Risk Map (June 2010)

Sources: Esri, HERE, Garmin, USGS, Intermap, INCREMENT P, NRCan, Esri Japan, METI, Esri China (Hong Kong), Esri Korea, Esri (Thailand), NGCC, (c) OpenStreetMap contributors, and the GIS User Community

NOTE: *The contents of this technical note are not to be used for advertising, publication, or promotional purposes. Citation of trade names does not constitute an official endorsement or approval of the use of such products.*