**Computing and Serving Environmental**
**Data Layers for Global Change Research using Google Earth Engine**
Proposal for a NCEAS Environmental Layer Working Group – Earth Engine partnership

*Introduction*

All of life on earth, whether crops, forests, insect pests or charismatic organisms like parrots and elephants exist in an environmental context. Indirectly, then, the environmental context is central to the distribution of biodiversity, the health and well-being of the human population that depends on organisms for food, clothing, ecosystem services (e.g. oxygen and water), health, and spiritual and recreational enjoyment. For a long time, biologists have been able to ignore the environmental context by largely treating it as a constant. Now, humans are changing the environmental context (climate, landscape, etc) so rapidly this is no longer viable. There is an urgent need for basic research and improved tools examining how this changing environmental context will impact organisms and in turn impact humans.

Over the last ten years, revolutions in earth observing systems and informatics have produced a stunningly well resolved view of climatology, terrain and land cover of the earth's surface that is in principle available to the global community via the Internet. In practice, this extraordinary new measurement of the environmental context of life is accessible to only a handful of scientists who have secondarily picked up skills in informatics, remote sensing and GIS. Repackaging these environmental data into formats that are accessible to the masses (scientists and the general public alike) presents an opportunity with profound implications for our ability to adapt and respond to unprecedented change. The few available products that at least partially meet this challenge demonstrate enormous demand and potential. The product most used among global change biologists is WorldClim, which according to Google Scholar has been cited in ~900 other papers in the six years since it was published and has been used in probably thousands more unpublished habitat modeling exercises. It has not, however, been updated since publication and contains a limited set of variables averaged over a fifty year time span (1950-2000) that do not capture the much more diverse and dynamic environmental information available today.

The dearth of easily usable environmental products means that global change biologists currently use products that rarely retain appropriate levels of complexity to encapsulate the often subtle spatial and temporal changes in climate, terrain and land cover critically important to all biodiversity and ecosystem processes. For example, the eastern and northern range boundaries of the saguaro cactus correspond with great precision to locations where the duration of freezes exceeds 12 hours, and Monarch butterflies die en masse when temperatures reach 42℃ for even one hour. Such critical thresholds are masked by data layers that commonly bin temperature data into monthly average values or provide only climatological means for minima and maxima. Although daily data that capture such extreme events are available, extracting the critical threshold values from them is not an option for most global change scientists given the expertise required and time-intensive nature of the work involved.

In contrast, readily available statistical techniques known as habitat models, niche models or distribution models now enable anyone to link the occurrence of species or other phenomena (e.g. disease prevalence) to their environmental context in a correlative fashion using gridded climatic and landscape data. Such approaches have become extremely popular. There thus is a compelling need to develop a set of simple to access and use, integrated environmental data products that both capture the spatial and temporal complexity of interest and disclose metadata about their derivation and associated uncertainties.

*Global Environmental Layers (GEL) working group and project*

As part of previous workshops (including the 'Map of Life' project, http://www.yale.edu/mapoflife), the PIs realized the tremendous need for a new set of reliable, community vetted, easy to disseminate and use, ecologically-tuned environmental data layers. Such ecologically tuned layers should improve the ecological realism of correlative distribution models, and thus their transferability in time and space, which is crucial when forecasting the effects of global change. PIs McGill, Jetz, McPherson and Guralnick were able to fund a National Center for Environmental Analysis and Synthesis (NCEAS) working group, bringing together experts in remote sensing, environmental statistics, species distribution modeling and informatics to ultimately produce such layers. The NCEAS working group first met in March 2010 in order to discuss the priority data layers, how they would be developed, and how they might be vetted in a species distribution modeling framework. We have just completed a "layers focused" second meeting as of Oct. 17th, 2010. At this second meeting, we began test runs of accumulating and using multiple set of data to assess different models and interpolation methods to produce the best possible climate, terrain and landcover products.

The objective of the project is to develop, integrate and serve to the public a comprehensive compilation of biologically relevant global environmental layers spanning the past fifty years. These layers will have a high temporal (e.g. weekly or monthly for each year) and spatial (<=1km) resolution and with minimal spatial error. The compilation should offer users continuously updated information and allow them to interactively calculate synthesized layers on-the-fly. The NCEAS working group members represent a well-developed community of users for such data (see Appendix I), have identified product requirements (see Appendix II), and have provided the scientific expertise to guide product development. In short, the social investment has already been made and we are now at the cusp of embarking on the technical investment.

*Use Cases*

*Farmers adapting to climate change.* A group of farmers in a tropical country with poor climate records would like to learn about the historical climatic conditions on their lands. The nearest weather station is 200 km away so cannot provide relevant data. Climate products fusing data from meteorological stations and satellite sensors produced by the Global Environmental Layers (GEL) project provide monthly weather averages for the past fifty years for every square kilometer of their country. Ongoing updates facilitated by Earth Engine's computational power allow for inclusion of the very latest information. The farmers identify through a map interface the pixel at the center of their lands, select the climatic variables they are interested in, are able

to download all the data and view a simple time-series graph. They find evidence that over the past fifty years the climate on their land has become drier and the ongoing trend suggests a long-term change in crops.

*Identifying weather-crime linkages.* A criminologist is seeking to link data on certain types of crime, collected worldwide over the past forty years, to climatic conditions at the point of their occurrence. Previous studies have shown that it is neither heat nor humidity per se but their interaction that drives crime waves. She uploads the points to the webpage, selects from existing derived variables such as AET (Actual EvapoTranspiration) or provides a formula to calculate a certain index, and then downloads the spatio-temporally specific climatic data for use in her analysis.

*Assessing the impact of coral bleaching on the sustainability of reef fisheries.* Anomalously warm water temperatures have been recognized as a major cause for coral bleaching. Coral bleaching occurs when coral organisms expel small algae known as zooxanthellae that usually live within the corals in a symbiotic relationship and can lead to the death of corals and ultimately the erosion of reef structure. Loss of structural complexity in turn affects fish community composition by rendering certain species prone to local extinction due to, for example, loss of shelter from predation. In the absence of reliable records on coral bleaching events, a coral reef biologist interested in quantifying sustainable levels of fishing uses Earth Engine to extract a medium-term index of reef exposure to anomalously high sea surface temperatures. She includes this index as an alternative explanatory variable in her analysis of how fishing effort affects fish biomass. She further uses data available on Earth Engine to verify that her study sites are comparable in terms of reef habitat and primary productivity.

*Monitoring pika populations under climate change.* Both heat stress and cold stress have been implicated in the decline of the American pika, a small rabbit-like animal that lives in alpine areas in Western North America. Extreme cold and warm events likely drive local extinctions of pikas. Pika researchers use Earth Engine to query for extended extreme cold and extreme warm events over a ten year period, in low and high precipitation years across montane and alpine areas in the Sierras, Rockies and Great Basin. Researchers can choose different severity and length of extreme events and automatically generate new outputs from Earth Engine. The same pika researchers also set up an alerting service such as a really simple syndication (RSS) feed to get updating reports of new such extreme events happening in montane and alpine areas where pikas occur.

*Targeting malaria intervention efforts*. Malaria is a mosquito-borne infectious disease that kills more than a million people each year. One of the first niche models ever developed was for disease bearing tse-tse flies, and niche models are now regularly used in predicting malaria risk. Tasked with developing a cost-effective malaria prevention program, public health officials in Laos work with a locally-based NGO to develop regional risk maps. Using Earth Engine, they develop a spatial query that links satellite-derived data on land cover, average monthly rainfall and temperature to the geographic area served by each of the country's clinics. The environmental data are then regressed against data on malaria case loads per clinic collected

over the previous decade. The resulting model is tested against the most recent data and found to be highly accurate. Blood serum samples collected by researchers working for the NGO indicate, however, that the prevalence of malaria varies widely even among villages served by the same clinic. To ensure that netting and other preventative tools first reach those in most need, the researchers develop finer-scale models for medium to high risk areas using high resolution terrain and soil moisture variables available from Earth Engine that better capture local-scale differences in mosquito breeding and disease transmission opportunities.

*The need and role for Earth Engine*

*We believe that Google Earth Engine will help overcome two key technical challenges that the Gobal Environmental Layers project faces. The first is the need for massive computing powers during the calculation, integration and analysis of layers. The second is in the delivery and dissemination of the data to the public. We describe each of these in more detail.*

*Calculation of layers*

The generation of 'smart' (ecologically-tuned) global environmental variables requires extensive calculations over petabytes of data (Figure 1). Specifically, we will need to produce two intermediate products, each many terabytes in size, which will then be further analyzed to provide summary products to end users. One intermediate product will contain daily climatic data for 50 years at a 1km scale across the globe, generated by merging a much larger set of satellite (MODIS) input data with data from meteorological stations. The second intermediate product will be a set of about 20 variables calculated worldwide at a 30m or 90m scale from the NASA SRTM (space shuttle radar measurements of land elevation). Both products should also be updated in near real-time as additional input data become available in the future.

Analyses on these intermediate products will allow us to generate a set of pre-computed data layers that exemplify the production of ecologically relevant environmental indices and that are available to users for immediate application. In an ideal world, users would moreover be able to build on these pre-computed layers by defining their own indices and having them calculated from the intermediate products in real time.

Several of these steps, especially the initial calculations on petabytes of data, the real-time updates and on-the-fly user formulated calculations are completely unfeasible on typical workstation type computers. The parallel computing power of Google Earth Engine represents an ideal solution to these challenges.

*Data Delivery and Dissemination*

The second major challenge we face is serving the data to end-users. We anticipate that users will want to interact with the product in three modes:

- Download tiles for use in advanced spatiotemporal calculation tools like Matlab, R, and ArcGIS
- Use a scroll-and-zoom interface (such as in existing Google Maps) to examine areas of interest on screen and then possibly download data matching the frames chosen

- Upload a list of geographic coordinates and download a matching list with the values of select of environmental variables at these points.

An additional challenge will be the temporal component. It is very likely that the GEL project could help extend Google Earth Engine's capabilities into the temporal domain. All of this needs to be done in a scalable, backed-up 24x7 environment which becomes a prohibitive environment to maintain and fund by the scientific community.

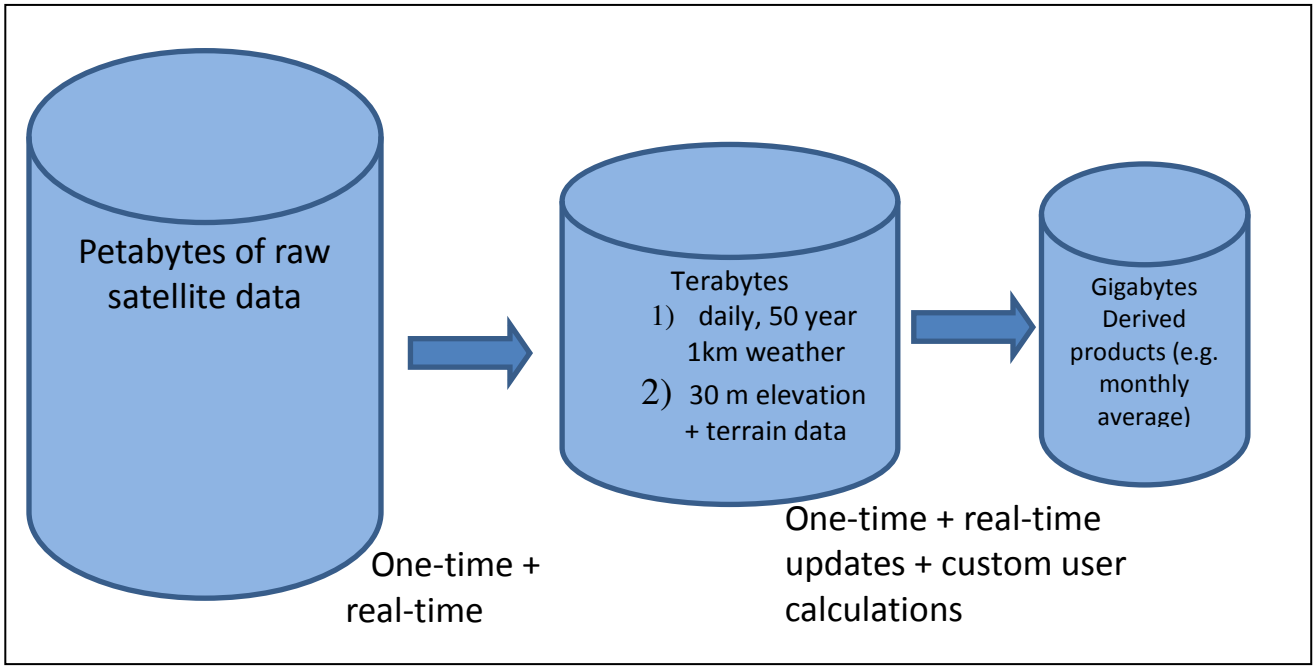


Figure 1. Compilations of summary products derived from in-situ and remote sensing data sources.

*Specific technical needs from Google Earth Engine and elsewhere*

Based on very preliminary discussions we anticipate that the following would be needed to successfully integrate GEL and Google Earth Engine (GEE) projects (not all of it necessarily provided by GEE – we are open to pursuing other funding sources where appropriate):

- Importing to GEE the appropriate raw data (see Appendix 2)
- Some training and support in the use of the GEE APIs to perform calculations
- Enhancement of the GEE to provide some of the machine learning tools commonly used by GEL (e.g. generalized linear models, regression trees)
- Possible extension of end-user interfaces to allow temporal viewing and selection (this could be a very nice feature for GEE in general – watching a time lapse overhead movie of deforestation for example)
- Exploration of facilities for automatic updating as new satellite data becomes available
- Java programming resources

## *Value Proposition Summary*

We anticipate that a partnership between the GEL project (currently manifested as working group funded by the National Center for Ecological Anlalysis and Synthesis) and the Google

Earth Engine team could be highly beneficial to both sides. Specifically, we would hope that the following values are perceived:
GEL provides:

- An institutionalized end user community representing the most commonly used tool for understanding and predicting the impacts of climate change on the biosphere (i.e. niche models)
- An assemblage of scientific expertise in the choice and development of data layers
- Scientific staff able to work on implementing the layers
- The opportunity to develop and serve data products of interest to a wide and diverse group of not only scientists but also the general public

GEE provides:

- Badly needed computing/data management resources
- A state-of-the-art user interface to the data
- Dissemination to the largest possible audience

## *Appendix I - Members of the NCEAS (National Center for Ecological Analysis and Synthesis) Global environmental layers working group*

*Active members*: Brian McGill, Univ. of Maine (lead PI); Jana McPherson, Calgary Zoological Society (Co-PI); Walter Jetz, Yale University (Co-PI); Robert Guralnick, University of Colorado at Boulder (Co-PI); Simon Ferrier, CSIRO; Jane Elith, University of Melbourne ; Steven Phillips, ATT Labs -Research; John Gallant, CSIRO; John Donoghue, University of Arizona; Florencia Sangermano, IDRISI, Clark University; Thiago Rangel, Universidade Federal de Goias; Stephanie Pau, NCEAS; Lauren Buckley, University of North Carolina; Tien Ming Lee, Yale; Julia Baum, NCEAS; Rodney Forster, CEFAS; Rob Hijmans, University of California at Davis

*Advisors*: Woody Turner, NASA; Ramkrishna Nemani, NASA; Kenneth Casey, NASA; David Foley, NOAA; Frank Muller-Karger, University of South Florida

## *Appendix II – List of planned Environmental Context of Organisms layers and required raw data inputs*

*Global Environmental Layers - Datasets and Algorithms*
Below we outline the products we envisage and list the datasets and algorithms needed to develop them. The goal is to produce global, gridded products at 1km resolution. This choice is a matter of practicality for users; datasets with higher than 1km resolution become increasingly intractable to download and use. If, however, Google Earth Engine can serve as a centralized workbench for environmental dataset generation and ultimately species distribution modeling, then some of these limitations disappear and there may be value to storing the data in the finest resolution available with aggregation happening on-the-fly as needed.

TERRAIN
Products (1km resolution):

- Worldwide DEM based on SRTM and ASTER
- Slope - steepness
- Aspect - exposure of surface to sun, wind and rain
- Catchment area (flow accumulation) - can be used to define channels, proximity to water etc but only in steeper areas
- Topographic wetness index - as a measure of accumulation of water and soil; not valid in flatter areas
- Multi-resolution valley bottom flatness - a surrogate for depositional zones of the landscape; also a good indicator of where topographic wetness index ceases to be meaningful
- Topographic position measure(s) - elevation relative to local mean, or within local range of elevations; frequently a useful ecological predictor
- Cliff map - as a specific unique habitat
- Solar insulation (monthly) - solar radiation for plant growth, effect on surface temperature; intimate link with water balance (radiation drives evaporation; soil water 'consumes' radiation by evaporation and moderates impact on surface temperature).
- Indices of spatial heterogeneity in the above layers as characterized by their mean, minimum, maximum, range, standard deviation, skew, spatial patchiness within each 1km grid cell

Raw data:

- Digital elevation model (DEM) at finest resolution available (3 second ~= 90 m, 1 second ~= 30m): SRTM, ASTER G-DEM
- Merge SRTM and ASTER to provide a global DEM (currently SRTM does not extend to the northern parts of Canada and Asia)

<u>CLIMATE</u>

Products (to be derived from an intermediate daily, 1km resolution product that fuses satellite-derived land surface temperature and precipitation data with climate station records):

*Basic Climate*

- Monthly climatology at 1km (30 second) resolution with monthly and annual minima, maxima and average summaries for:
    - Air temperature 2000 - 2009
    - Air temperature 1950- 2009
    - Precipitation 2000 - 2009
    - Precipitation 1950- 2009

*Extreme events*

- Temperature of the coldest day of an average year
- Number of frost days or frost free days
- Frost duration
- Date of first frost
- Date of last frost

*BioAg variables*

- Growing Degree Days
- Actual Evapotranspiration rate

- Potential Evapotranspiration rate
- Bouwmans Soil Water Holding Capacity.
- Thornthwaite water balance
- Palmer Drought Severity Index
- Vapor pressure deficit

Raw Data:

*Land Surface temperature*

- MOD11A1 product: resolution 1 km // tile size: 1200 km2 tiles // coverage: global // temporal dimension: 2000 to 2010
- MODIS daily day time land surface temperature bands at 1km resolution
- MODIS daily night time land surface temperature bands at 1km resolution
- MODIS daily day time quality control bands
- MODIS daily night time quality control bands
- MODIS clear day coverage (% of pixel covered by clouds)
- MODIS clear day coverage (% of pixel covered by clouds)

*Precipitation*

- TRMM - tropical rainfall measuring mission:
    - Monthly (3B43): resolution: 0.25deg // tile size: global // coverage: 50N - 50S // temporal dimension: 1998 - 2010
    - 3 Hours (3B42) : resolution: 0.25deg // tile size: global // coverage: 50N - 50S // temporal dimension: 1998 - 2010

*Climate station data*

- Global historical climatology network
- USHCN (US Historical Climate Network)

<u>LANDCOVER</u>

Products:

- 'naturalness' index
- proportional coverage per grid cell for each of 23 global land cover classes

Raw data:

- Globcover version 2.2 at 300m (10 second) resolution
- Global forest canopy height (500m)
- Impervious surfaces (500m)
- Urban extent (500m)
- Percent tree cover (500m)

<u>MARINE</u>

Products:

- actual values and climatology values (averaged over a minimum of 10 years) for sea surface temperature (SST), turbidity and primary productivity at weekly and monthly intervals and a spatial

resolution of 1km (or if need be 4km), blending data gathered by multiple satellite and in situ sensors where possible

- coral bleaching disturbance indices at 1km (4km) resolution: weekly SST anomalies, counts of weekly anomalies by month, year and over the total available time series, minimum and mean return intervals of anomaly events likely to induce bleaching
- light starvation indices at 1km (4km) resolution: average (based on climatology) and annual maxima of turbidity for periods of one, two, three and four consecutive weeks, minimum and mean return intervals of (ideally user-defined) upper threshold turbidity values
- visual exposure indices at 1km (4km) resolution: average (based on climatology) and annual minima of turbidity for periods of one, two, three and four consecutive weeks, minimum and mean return intervals of (ideally user-defined) lower threshold turbidity values
- resource starvation indices at 1km (4km) resolution: average (based on climatology) and annual minima of primary productivity for periods of one, two, three and four consecutive weeks, minimum and mean return intervals of (ideally user-defined) lower threshold productivity values
- coral reef habitat indicators at 1km resolution: dominant habitat type, percentage cover per type and habitat diversity
- terrain features at 1km: bathymetry, topographic position

Raw Data:

*Sea surface temperature*

- single sensor SST products from MODIS on Aqua and Terra (1km resolution, 2000-ongoing), AATSR on ENVISAT (1km, 2002-ongoing), and AVHRR on NOAA-Series satellites (4km, 1981-ongoing).
- In situ SST measurements collected by the Global Ocean Data Assimilation Experiment (GODAE http://www.godae.org/).
- Blended 1km daily SST from G1SST (2008-ongoing, (http://ourocean.jpl.nasa.gov/SST/)

*Turbidity*

- Single sensor diffuse attenuation coefficient at 490nm from MODIS on Aqua and Terra (2000-2010), OCTS (1996-2010), SeaWiFS (1997-2010), MERIS (2002-2010)
- Blended daily diffuse attenuation coefficient at 490nm (http://catalogue.myocean.eu.org/)
- Blended daily secchi disk depth (http://catalogue.myocean.eu.org/)

*Productivity*

- Chlorophyll products from Aqua-MODIS (2000-2010), SeaWiFS (1997-2010) and possibly other sensors
- Measures of photosynthetically active radiation from Aqua-MODIS (2000-2010), SeaWiFS (1997-2010) and possibly other sensors

*Reef Habitat*

- Vector data layers from the Millennium Coral Reef Mapping Project (http://www.imars.usf.edu/MC/index.html)
- Binary 1km resolution reef map from the World Resources Institute

*Bathymetry*

- SRTM30_PLUS