A Flux Atlas for Representativeness and Statistical Extrapolation of the AmeriFlux Network

William W. Hargrove
Environmental Science Division
Oak Ridge National Laboratory
Building 1507, Room 211, M.S. 6407
Oak Ridge, TN 37831–6407
(865) 241–2748
hnw@fire.esd.ornl.gov

Forrest M. Hoffman
Computer Science & Mathematics Division
Oak Ridge National Laboratory
Building 5600, Room B109, M.S. 6008
Oak Ridge, TN 37831–6008
(865) 576-7680
forrest@climate.ornl.gov

April 28, 2004

Contents

List of Figures

List of Tables

Abstract

The ecoregion concept has proven useful as a way to generalize the complex multivariate aspects of environmental conditions for plants and animals. Grouping such complex environmental combinations into a series of relatively homogeneous areas on a map helps ecologists to better understand the complex factors affecting the distribution of organisms across the landscape. Ecoregions have both theoretical and practical uses, and often form the basis for sampling designs and conservation strategies.

Here we borrow the utility of the ecoregion concept and apply it to the problem of sampling and estimating the amount and direction of carbon flux from terrestrial ecosystems. Unlike hand-drawn ecoregions produced by human experts, however, we employ a multivariate clustering procedure using a parallel supercomputer to produce a spectrum of homogeneous carbon flux regions. Each of these flux ecoregions is created statistically on the basis of a number of input maps so that ecosystems within a single flux ecoregion would be expected, from first principles, to have about the same magnitude and direction of carbon flux. We do not try (yet!) to estimate the magnitude or direction of carbon flux for each homogeneous flux ecoregion. Instead, our strategy is to statistically pre-stratify carbon flux environments across the conterminous United States as a way to simplify subsequent carbon flux sampling, measurement, and estimation.

As many as 30 maps were used as input to the multivariate clustering procedure to produce a spectrum of homogeneous carbon flux ecoregions for the conterminous United States. The spectrum represents different combinations of input maps and seasonality. Most input maps were created specifically for this project, and were themselves created of many individual components. Input maps were constructed to minimize cross-correlation with other inputs and to maximize the content of unique information relevant for carbon flux processes. For example, an elevation layer was not used. Instead, we included degree-day heat and cold sums, diurnal temperature differences, and precipitation totals, since the effects of elevation on vegetation are probably mediated through temperature and precipitation.

A number of investigators are trying to scale flux tower measurements from the bottom up to represent larger geographic regions. The flux ecoregion approach, on the other hand, relies on remotely sensed data to scale flux tower measurements to continental scales in a top-down way. It may be easier to obtain continental-scale flux fields to satisfy the goals set by the North American Carbon Program (NACP) and the AmeriFlux Program using top-down approaches, although both strategies will be needed for verification and testing.

If funded, a suite of AmeriFlux-specific results will follow in the second half of the project, based on the statistical foundation provided by this flux atlas. Although we use map colors to help visualize multivariate similarities, our statistical techniques are fully quantitative, and similarities and differences between different carbon flux environments can be expressed numerically. We will use the quantitative similarity of the suite of ecosystem characteristics that are relevant to carbon fluxes to extrapolate from existing flux measurements to estimate fluxes within ecoregions where fluxes are unmeasured. We hope to statistically model the multidimensional relationship of carbon flux with respect to each flux ecoregion characteristic, and then modify the expected flux according to the quantified differences in the flux environments between the unmeasured flux ecoregion and the most similar flux ecoregion containing an AmeriFlux tower. This quantitative ecoregion-based, top-down approach to stratifying carbon flux and

then estimating flux from existing measurements, if successful, may be the fastest way to help fulfill the NACP and AmeriFlux Program goals of seasonally mapping regions of sources and sinks of carbon within the North American continent.

Chapter 1

Introduction

The North American Carbon Program (NACP) Plan for carbon cycle research is focused on measuring and understanding the sources and sinks of CO_2 in North America and adjacent ocean regions, and is an important component of both the interagency U.S. Carbon Cycle Science Program and the new U.S. Climate Change Science Program. Although individual towers in the AmeriFlux network give important information about carbon flux in the immediate area surrounding each tower, there are insufficient numbers of towers to provide the regional and continental scale information on sources and sinks desired by NACP. It is unlikely that enough towers can be added to the network to make the density of measurements adequate for spatially continuous continental resolution within the timeframe required by NACP. A method for extrapolating spatially localized measurements into a continuous field across the continent is needed.

This project will ultimately determine the degree to which the existing network of carbon eddy flux towers within the AmeriFlux network are representative of flux environments across the continental United States (CONUS). Better data quality maximize the chances of success within the CONUS; if successful, the process can be expanded to the North American continent. This representativeness information can be used to determine how many additional towers will be required, and where additional towers should be placed. In addition, the uniqueness of each existing tower to the AmeriFlux network will be calculated. More importantly, this work will provide a statistical basis for quantitatively determining the similarity of flux environments in two different locations. If flux has been measured in one of these environments, it may be possible to "adjust" the tower measurement, using the quantitative differences between the flux environments, to obtain a reasonable estimate of flux in the unmeasured location.

In the first phase of this project, we have statistically created a series of flux-relevant ecoregions which divide the conterminous U.S. into a set of areas within which the carbon flux from terrestrial ecosystems is expected to be relatively uniform and homogeneous. Starting with digital geographic information system (GIS) layers of factors deemed important in regulating carbon fixation and loss from terrestrial ecosystems, we have created a set of multivariate descriptors which characterize the environment in each map cell. Then, we used a k-means clustering procedure to classify each map cell into a particular group whose cells have sufficiently similar environments. The classification procedure is iterative and converges on a final assignment solution when fewer than a specified number of cells change from the last assignment.

Each environmental descriptor is itself a map, and most maps were developed specifically for this project. These maps are at 1km^2 resolution over the conterminous United States, and consist of nearly eight million cells each. Because there are as many as 30 environmental descriptor maps, each with nearly eight million cells, it is necessary to perform the clustering process on a parallel supercomputer.

Because the process is quantitative, the similarity of a selected flux-ecoregion to every other ecoregion in the map can be calculated. A map can be produced that shows the degree of similarity to the chosen flux-ecoregion as a series of gray shades. By sequentially selecting the regions currently containing an AmeriFlux tower, a map of the geographic area which is represented by that flux tower can be produced.

This project began at the end of FY2003 and represents the first phase of work proposed in a solicited proposal to the U.S. Department of Energy (DOE). This report presents maps of homogeneous flux-relevant ecoregions and describes how we obtained them. Based on the reviews of this report, a decision will be made on the funding requested for the second phase of the proposed research.

The second phase of the project will utilize these flux-relevant ecoregions to determine the overall representativeness of the existing AmeriFlux network, calculate uniqueness values for each tower, and determine the most favorable locations for the next five towers. Additionally, we will utilize the quantified similarities of flux-ecoregions having AmeriFlux towers to those without towers in order to estimate **net ecosystem exchange (NEE)** across the U.S. By using the similarity values as weight coefficients, it may be possible to "adjust" flux measurements within measured flux-ecoregions by how similar their environments are to flux-regions which have no measurements.

Representativeness analyses are being conducted at the ecoregion scale at 1 square kilometer resolution. Results obtained will be with respect to the particular ecosystem characteristics upon which the flux-ecoregions were based. If a characteristic which is overridingly important for carbon flux has not been included, or if a characteristic which is not related to flux has been included, the relevance of the ecoregions which are produced will be decreased. Unusual local site conditions such as managed ecosystems, disturbances, or topographic peaks or bowls may make particular tower locations less representative of the square kilometer within which they sit. However, this resolution is probably a good match with the "fetch" of an individual flux tower when integrated over time.

1.1 The Multivariate Geographic Clustering (MGC) Method

Hargrove and Hoffman have developed a method based on a multivariate statistical approach to divide any map, based on a set of multivariate characteristics, into a set of regions which are relatively homogeneous with respect to those conditions (Hargrove and Hoffman, 2004). The number of regions which are produced is under user control. Once produced, any two regions taken from the map will have roughly equivalent within-region heterogeneity with regard to the combinations of characteristics used as input. The process utilizes information contained within the variance structure of the input data, and creates regions such that the variance is equally divided among all of the resulting regions. Regions that are produced may be spatially disjoint; for example, environments on geographically distant mountaintops, if similar enough, will be classified into the same flux-ecoregion.

Because the number of characteristics considered can be very large, and because the process operates on a cell-by-cell basis on high-resolution maps, a parallel supercomputer is needed to produce such regionalizations (Hargrove et al., 2001). The non-hierarchical algorithm consists of a reversible transformation between two realms, one in two-dimensional geographic space, and one in multi-dimensional data space. Normalized variable values from each map cell are used as coordinates to plot each map cell in an environmental space with as many axes as there are multivariate environmental dimensions. Because the plotted location of map cells in data space pinpoints the combination of environmental variables within that map cell, two map cells which are plotted close to one another in data space will have similar mixtures of environmental conditions, and are likely to be classified into the same ecoregion cluster. Thus, similarity is coded as the separation distance in this data space. We use an iterative k-means clustering algorithm (Hartigan, 1975), which begins with a user-specified number of ecoregion clusters, k, into which the map cells are to be grouped. All map cells are examined sequentially to find the most widely separated set of cells which will provide k initial "seed" **centroids**, one for each of the desired k cluster groups. In a single iteration of the clustering procedure, each map cell is assigned to the closest (*i.e.*, environmentally most similar) centroid in terms of Euclidean distance. At the end of the iteration, once all map cells are assigned to a centroid, the coordinates of all map cells within each group are averaged to produce a new, adjusted centroid for each cluster, and another iteration of assigning map cells to these new centroids begins. The iterative process of classifying map cells and adjusting centroid locations continues until fewer than a pre-determined number of map cells change cluster assignments during an iteration. After the process has converged on a particular grouping scheme, the k ecoregions have been statistically defined.

Once cluster assignments have stabilized, map cells are re-assembled in geographic space, retaining their ecoregion classifications. Although geographic coordinates are not used directly in the classification, ecoregions tend to be geographically cohesive due to the spatial autocorrelation which is usually present in the environmental data. Because of the Euclidean distance assignment method, the k-means clustering algorithm tends to fit globular clusters of equal size in data space. Thus, all large ecoregions share a similar upper limit on within-group variance, and have a similar maximum radius around each centroid. This uniform heterogeneity across ecoregions prevents the creation of side-by-side ecoregions that have vastly different within-region variance.

We call this empirical process **Multivariate Geographic Clustering (MGC)**. We have implemented it in a parallel algorithm written in the C language using the Message Passing Interface (MPI). The code is dynamically load-balancing, fault-tolerant, and performs both initial seed-finding and iterative cluster assignment in parallel. Individual nodes independently classify subsets of cells, then combine results at the end of an iteration. As a result, the quantitative ecoregionalization algorithm can be used to analyze very large datasets quickly (Hoffman and Hargrove, 1999).

Ecoregions determined using MGC are self-describing in that the coordinates of the final centroids quantitatively define the synoptic conditions for each ecoregion. Nominative multivariate conditions within a particular ecoregion are described by the N coordinates of its centroid. The technique is parametric only in that means are used to calculate each new centroid. Rather than imposing a pre-conceived external grouping upon the map, the variance structure present in the environmental conditions is used, allowing an unbiased, uniform classification structure to emerge from the data.

Initially, we performed a number of empirical regionalizations for the conterminous United States at 1km^2 resolution, dividing the U.S. into as many as 3000 distinct ecoregions (Hargrove and Luxmoore, 1998). We included nine characteristics from three categories—elevation, edaphic factors, and climatic factors. The edaphic factors were 1) plant-available water capacity, 2) soil organic matter, 3) total Kjeldahl soil nitrogen, and 4) depth to seasonally-high water table. The climatic factors were 1) mean precipitation during the growing season, 2) mean solar insolation during the growing season, 3) degree-day heat sum

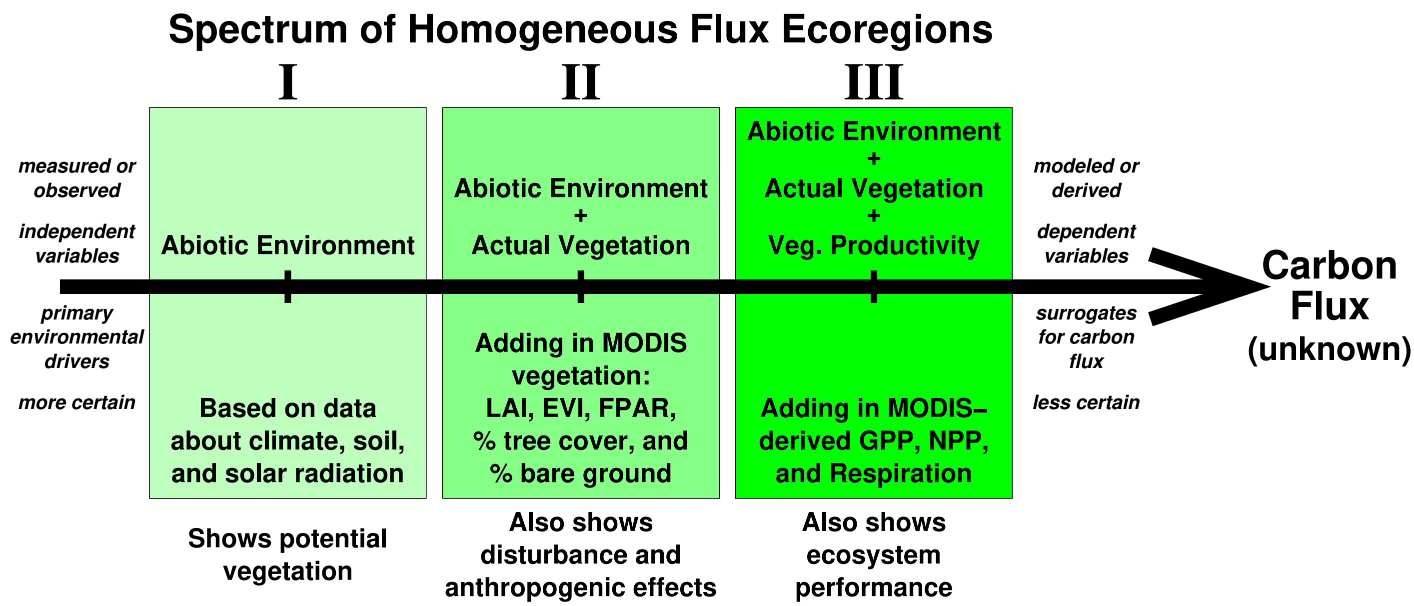

Figure 1.1: The spectrum of Homogeneous Flux Ecoregions

during the growing season, and 4) degree-day cold sum during the non-growing season. The growing season was defined by the frost-free period between mean day of first and last frost each year. A map for each of these characteristics was generated from best-available data at a $1km^2$ resolution for input into the clustering process. Each of the input maps contained more than 7.8 million cells. Such map, data, and ecoregion resolution surpasses that usually accomplished by ecoregion experts using qualitative methods. These maps appear to capture the ecological relationships among the nine input variables. More recently, we have added new variables and divided the U.S. into as many as 5000 ecoregions. Twenty-five environmental factors, including elevation, mean and extremes of annual temperature, mean monthly precipitation, soil nitrogen, organic matter, water capacity, frost-free days, soil bulk density and depth, and solar aspect and insolation were included (Hargrove and Hoffman, 2004).

1.2 Conceptual Approach to Mapping Flux Ecoregions

The conceptual basis for development of the flux-relevant ecoregions is a spectrum or gradient starting from primary ecological forcing factors on the left, and building additively toward the right (Figure 1.1). Actual carbon flux, or NEE, forms the extreme right side of this continuum. NEE is unknown and is not directly measured for most ecosystems in the CONUS, and is what we are trying here to design a network to measure. If we knew the continuous map of NEE, we could design a perfect network to monitor and sample it. However, we wouldn't need to measure it, since we would know it already! So the extreme right hand side of this continuum is ultimate-but-unknown NEE.

Our approach has been to start with primary factors on the left, and then add further information about the primary characteristics of the existing vegetation, and finally to also add information about simulated surrogates of ecosystem performance. In this approach, then, we start with the simple, and add factors which are increasingly complex, building toward the unknown, NEE itself. Because uncertainty also increases from left to right along this continuum, this step-wise approach to regionalization initially produces oversimplified, but more certain initial flux-relevant ecoregions, but ultimately produces more sophisticated, but less certain flux-ecoregions, all based on best-available data.

The left side of Figure 1.1 represents climatic, edaphic, and physiographic factors which might be considered, *a priori*, to impose limits on the amount and direction of carbon flux for a particular ecosystem. We will produce a set of regionalizations based solely on these primary drivers. However, consideration of primary forcing factors alone will produce ecoregions which reflect the carbon flux associated with **potential vegetation**. Anthropogenic effects, disturbance regimes, and history may have altered or reset the potential vegetation. The middle of this conceptual spectrum adds information about the characteristics of the vegetation that is actually existing in this ecosystem at this time. Finally, in a third set of flux regionalizations, we will add information about ecosystem gross primary production and respiration. These ecosystem activity values are not measurements, but are simulated based on measurements, and can be thought of as surrogates for the unknown carbon flux itself. Thus, we will create three sets of regionalizations, each time adding factors used as input, and increasing in sophistication (Figure 1.1).

Each of the input maps had to be developed before they could be used. These are valuable products in themselves. Each was constructed from the best-available data at a 1km^2 resolution. Most of the input maps are themselves synthetic aggregate products which were produced from map series through time, and were designed to summarize or integrate the most important aspects of that raw map series into a single component which is both immediately important to vegetation and independent of information contained in other input layers. For example, we did *not* include elevation, since elevation effects are confounded with (and probably act through) differences in temperature and precipitation. Variable selection represents a compromise between desirability and availability.

Remote sensing provides the only practical source of data which can be used to describe the characteristics of existing vegetation for the middle of the conceptual continuum. We are using data from NASA's new MODIS sensor, which is now operating on both the Terra and Aqua orbital platforms. Only MODIS Terra data were used here, since the Aqua platform has not been up long enough to produce a substantial data record. MODIS provides estimates of **Leaf Area Index (LAI)**, **Fraction of Photosynthetically Active Radiation (FPAR)**, and an **Enhanced Vegetation Index (EVI)** at 1km^2 resolution. EVI is provided once every 16 days, while LAI and FPAR are available every eight days. The **Continuous Vegetation Fields** product provides estimates of **percent tree cover** and **percent bare ground** in every cell across the CONUS.

MODIS was also used to provide indices of ecosystem performance for the third set of flux regionalizations. **Net Photosynthesis (PSNNET)** and **Gross Primary Production (GPP)** are produced every eight days. We have taken the eight-day difference of PSNNET from GPP as a **respiration index (RI)**, which we believe is a more relevant measure of ecosystem performance. The **Net Primary Productivity (NPP)** product is also available for the last two years. MODIS **Land Surface Temperature (LST)**, during both the day and the night, will be used in the initial flux regionalization using only environmental forcing factors. As the temperature of the vegetation canopy or soil surface, LST should be more meaningful than just the temperature of the ambient air.

1.3 Separation into Local Growing and Non-growing Seasons

We are using a consistent approach to integrate raw time series maps into input layers by dividing the annual cycle at every cell into a growing season and a non-growing season. The growing season is defined by the frost-free period between the median day of the last frost of the spring and the median day of the first frost of the fall, as measured over the 30-year climate normals period (1960–1990). We used these dates to divide the synoptic year into two temperature-defined periods customized for each cell location in the CONUS. These two seasons form a common basis for dividing many of the input variables into periods that should be relevant for vegetation growing across the CONUS.

All variables other than edaphic factors were integrated over the growing and non-growing seasons into two summary maps. For example, a heat sum was constructed from daytime LST over the local growing season, and a cold sum was constructed from nighttime LST over the non-growing season. Similarly, total precipitation was summed over the growing and non-growing seasons. Because these layers contain information about precipitation totals, but not the temporal distribution of precipitation events, another layer was included which counts number of measurable precipitation events during the growing and non-growing seasons.

1.4 Notation for Alternative Flux Ecoregions

This work has produced nine separate sets of flux ecoregions for the conterminous U.S. (three types of flux ecoregions $\times$ three seasons for each type). We have adopted a shorthand notation for distinguishing these alternative flux regionalizations which is shown in Figure 1.2. We use Roman numerals I, II, and III to refer to the three different stages along the conceptual continuum shown in Figure 1.1 from primary forcing factors to derived ecosystem flux surrogates. Within each of these, we use Arabic A, B, and C to refer to annual, growing season, and non-growing season divisions, respectively. Thus, "IIB" refers to the flux ecoregions created using abiotic environment and actual vegetation factors which were integrated over the growing season. We use this notation throughout the remainder of this report to refer to particular alternative flux ecoregions.

The type "A" annual regionalizations contain both growing season and non-growing season layers rather than a single annual layer, since the division into these two temperature-defined seasons is still a more relevant form than simply integrating across the entire year. This means that the "A" versions of each flux regionalization contain more input maps than the "B" and "C" versions (Figure 1.2).

Table 1.1 shows the list of input maps used to produce each set of flux ecoregions. We included 18 layers of primary forcing factors in developing the first set of flux regions (Series I). To these we added eight MODIS-derived layers from each of the two seasons to produce the regionalizations which consider existing vegetation (Series II). Finally, another four layers comprised of composite MODIS-derived flux surrogates were added to produce the most sophisticated flux regionalizations (Series III). Most of these input layers represent valuable products in themselves. But these groundbreaking products represent our starting point, not the finished product for this task.

	Abiotic Environment	Abiotic Environment + Actual Vegetation	Abiotic Environment + Actual Vegetation + Vegetation Productivity
Annual (growing and nongrowing)	IA 18 input maps	IIA 26 input maps	IIIA 30 input maps
Local Growing Season	IB 12 input maps	IIB 17 input maps	IIIB 19 input maps
Local Non–growing Season	IC 12 input maps	IIC 17 input maps	IIIC 19 input maps

Figure 1.2: Shorthand notation for distinguishing alternative flux ecoregionalizations

Map Layer or Variable Name	**IA**	**IB**	**IC**	**IIA**	**IIB**	**IIC**	**IIIA**	**IIIB**	**IIIC**
Degree-days heat sum above 42°F from daytime land surface temperature during the local growing season	✓	✓		✓	✓		✓	✓	
Degree-days cold sum below 42°F from nighttime land surface temperature during the local non-growing season	✓		✓	✓		✓	✓		✓
Number of days above 90°F during the local growing season	✓	✓		✓	✓		✓	✓	
Number of days below 32°F during the local non-growing season	✓		✓	✓		✓	✓		✓
95th percentile of maximum diurnal surface temperature difference during the local growing season	✓	✓		✓	✓		✓	✓	
95th percentile of maximum diurnal surface temperature difference during the local non-growing season	✓		✓	✓		✓	✓		✓
Precipitation sum during the local growing season	✓	✓		✓	✓		✓	✓	
Precipitation sum during the local non-growing season	✓		✓	✓		✓	✓		✓
Number of days with measurable precipitation during the local growing season	✓	✓		✓	✓		✓	✓	
Number of days with measurable precipitation during the local non-growing season	✓		✓	✓		✓	✓		✓
Depth of mineral soil	✓	✓	✓	✓	✓	✓	✓	✓	✓
Depth to water table	✓	✓	✓	✓	✓	✓	✓	✓	✓
Soil Kjeldahl nitrogen to 50 cm depth	✓	✓	✓	✓	✓	✓	✓	✓	✓
Soil organic matter to 50 cm depth	✓	✓	✓	✓	✓	✓	✓	✓	✓
Soil plant-available water holding capacity to 1.5 m	✓	✓	✓	✓	✓	✓	✓	✓	✓
Compound Topographic Index (CTI)	✓	✓	✓	✓	✓	✓	✓	✓	✓
Total solar insolation during the local growing season, including clouds, aerosols, slope and aspect physiography	✓	✓		✓	✓		✓	✓	
Total solar insolation during the local non-growing season, including clouds, aerosols, slope and aspect physiography	✓		✓	✓		✓	✓		✓
Enhanced Vegetation Index (EVI) integrated over the local growing season				✓	✓		✓	✓	
Enhanced Vegetation Index (EVI) integrated over the local non-growing season				✓		✓	✓		✓
Fraction of Photosynthetically Active Radiation (FPAR) absorbed by vegetation integrated over the local growing season				✓	✓		✓	✓	
Fraction of Photosynthetically Active Radiation (FPAR) absorbed by vegetation integrated over the local non-growing season				✓		✓	✓		✓
Leaf Area Index (LAI) integrated over the local growing season				✓	✓		✓	✓	
Leaf Area Index (LAI) integrated over the local non-growing season				✓		✓	✓		✓
Percent tree cover				✓	✓	✓	✓	✓	✓
Percent bare cover				✓	✓	✓	✓	✓	✓
Gross Primary Production (GPP) integrated over the local growing season							✓	✓	
Gross Primary Production (GPP) integrated over the local non-growing season							✓		✓
Respiration Index (RI) integrated over the local growing season							✓	✓	
Respiration Index (RI) integrated over the local non-growing season							✓		✓

Table 1.1: Map Layers Used in Multivariate Geographic Clustering

1.5 Development of Input Maps

We have processed every MODIS granule over the United States, for every eight-day interval, for every year, for 2001, 2002, and 2003 for almost every MODIS product from the Terra platform. There are 14 granules for each of the 46 periods for each of the three years for most MODIS products. We are using Version 4 or later for all of these products (for PSNNET and GPP, we are using version 4.5, direct from Dr. Steve Running's lab). Just acquiring all of the primary MODIS files in HDF format (14 granules per date $\times$ 46 dates per year $\times$ three years $\times$ five MODIS product families = 9660 primary files) via ftp was an accomplishment. We wrote custom scripts to handle each type of MODIS product, mosaic the granules together, change them into GIS layers, and reproject them into the **Lambert Azimuthal Equal-Area (LAEA)** projection for the CONUS, using a 6,370,997m radius for the Earth and centered on latitude 45 and longitude -100.

MODIS quality control is provided by flags in bit fields of an image which is separate from the primary image. Unfortunately, there is little consistency among various MODIS products in how these flags are to be read. Furthermore, there is as yet little agreement among the MODIS community about which combinations of these flags produce the best quality assurance. We checked our interpretations of the MODIS quality assurance flags with Dr. Steve Running, who believes that we are using these flags appropriately for the carbon flux regionalization task. We masked most images with the quality assurance flags, using only data which are of sufficiently high quality.

MODIS images are always taken on the same Julian dates each year, usually every eight days, but every 16 days for vegetation indices, producing either 46 or 23 images per year. Because we processed three full years of MODIS images (2001–2003) for most products, we have three separate images for each date. Since our goal is to develop full coverages for a single average or synoptic year, we composited the three separate images together to produce a single, more complete image for each Julian date. For most products, this composite is formed by averaging the three available year images for that Julian day.

1.5.1 Filling Gaps in Images

Even after compositing, holes exist in the composite images due to the QA flags, clouds, snow, etc. We developed methods for filling in such gaps in the images. Some MODIS products like LAI/FPAR also have internal flags for lakes, urban areas, perennial ice, swamps, etc. We set such areas to zero.

We developed a mathematical procedure, based on the values in each cell through time, to perform a temporal interpolation to fill gaps one cell at a time. A spline with tension was fitted through the temporal sequence of values in each cell. In order to pre- and post-condition the spline curve, we duplicated the 46/23 values in the synoptic year, and added an extra copy of the year onto the front and back, so that the spline was actually developed to fit a three-year temporal sequence. Interpolated values taken from the middle year were used to fill the gaps to avoid edge effects.

The spline procedure has an adjustable tension, which acts to draw the curve back to trend. To avoid overshoots, we slightly expanded the range of the actual data values, and then fitted the spline repeatedly, starting with the lowest possible tension. If the interpolated values exceeded the new range of the data, we increased the tension and re-calculated the spline. In this way, the interpolated values were always near the range of the actual image values, and were obtained using the lowest possible tension setting.

Because of this limited overshoot, exclusively positive values may be caused to occasionally go slightly negative in a few cells. If the totals over time are still negative, we reset these few cells back to zero.

We believe that this temporal interpolation using splines is the most sophisticated procedure currently being used to fill gaps in MODIS products. Since most MODIS parameters are temporally autocorrelated, making use of the temporal sequence of information at a cell location is likely to be better than relying on spatial neighbors for interpolation. While some researchers are using temporal linear interpolation, we have seen many examples in temporal sequences of MODIS data where such linear models seem less appropriate than a spline curve.

Nevertheless, there are occasionally map cells which have three or fewer values over the entire 46/23 image synoptic year. A temporal spline would not likely have sufficient information to appropriately fill data gaps under such data-poor circumstances. In these cases, we switch to a spatial interpolation method to fill the holes in the maps. A weighted circular adaptive spatial filter with a radius of 3km replaces the focal center cell with the weighted mean of all data-containing cells within the current range of the filter. The filter is used in multiple passes until all remaining holes are filled.

1.5.2 Integration of Input Variables Through a Season

To increase relevance for vegetation and carbon flux, we integrated the area under the time curve by growing and non-growing season. Photosynthesis and respiration are proportional, for example, to the area of the leaf crop as indicated by LAI times the number of days that the vegetation had that leaf crop (Prince, 1991; Rasmussen, 1997). Carbon flux, in turn, should be related to photosynthesis and respiration. A similar conceptual argument can be made for integrating EVI, FPAR, PSNNET, and GPP by season.

Integration by season helps to distinguish not only evergreen from deciduous vegetation, but also mixtures of each type. We are using the spline with tension to turn the discrete values at each of the 46/23 image periods into a smooth curve with a unique value for each of the 365 days of the year. These daily values are then summed in the integration step to produce three maps: an annual integration, a growing season integration, and a non-growing season integration.

The growing and non-growing season maps sum to what would be full annual values (Figure 1.3). Sometimes this temperature-defined seasonal split causes surprising features in the map. For example, the non-growing season total solar insolation map shown in Figure 1.3 makes it appear that the southern tip of Florida does not receive much sun. This is because the southern tip of Florida has no non-growing season over which to integrate insolation. The summed map at the bottom has the appearance that we might intuitively expect. While such annual maps look more understandable, seasonally divided maps are likely to be more relevant for plants, and are a better match for the time-scales over which we will be using the flux ecoregions to estimate flux measurements.

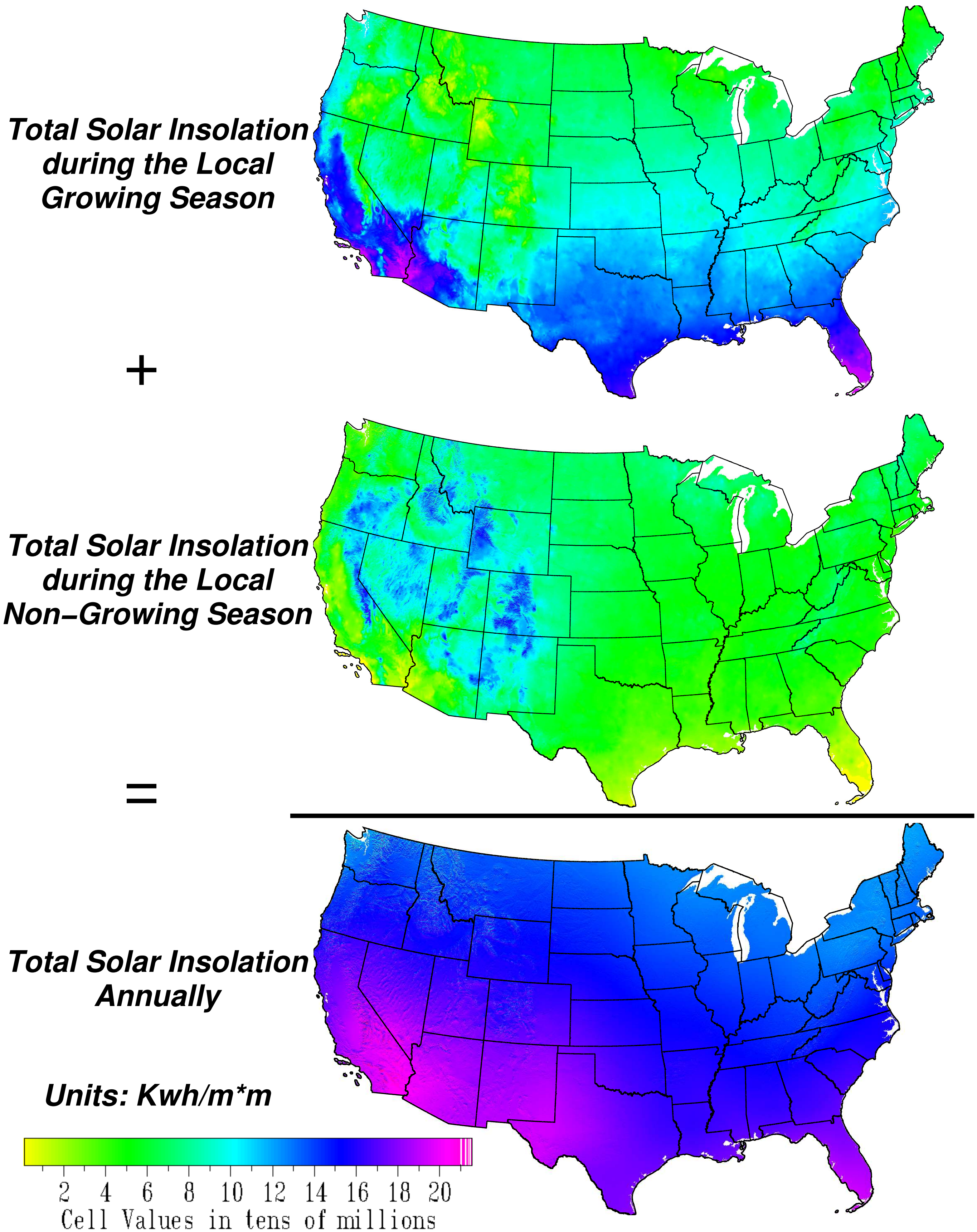

Figure 1.3: Annual Total Solar Insolation

While dividing these maps by the number of days in the growing season (or simply not time-integrating) would produce an instantaneous map which might be more intuitive, a map which is integrated through time is more useful for estimating seasonal carbon flux. For example, two different flux ecoregions might, over the course of a growing season, absorb roughly equal amounts of carbon, but one might do it in a short spike of fixation whereas the other does it with a small but constant fixation rate. Because we are interested in source/sink terms over seasonal intervals, and because we will be integrating measurements at flux towers over seasonal time scales, season-integrated maps are more relevant for the creation of these flux ecoregions.

We used a variation of the spline procedure to calculate 95^{th} percentile of maximum diurnal temperature difference in the growing and non-growing season. First, a spline was fitted to the 46 MODIS daytime Land Surface Temperature (LST) values, and was then used to estimate 365 daily values of daytime LST. This produced 365 maps of the CONUS, each with 7.8 million cells. Next, the splining procedure was repeated for MODIS nighttime LST values, producing another set of 365 CONUS maps. Finally, these maps were subtracted from each other, restricted cell-by-cell to the spatially unique growing and non-growing season, and the results sorted by cell to find the 95^{th} percentile value. We used the 95^{th} percentile value rather than the maximum diurnal difference to avoid any problems that could be caused by mapping anomalies.

In addition to MODIS products, we included layers that we have developed ourselves, along with a few that we have purchased. As with the MODIS products, most of these raw layers were not included as inputs to the MGC process directly. Rather, they were used as components in the construction of second-order products. These second-order products, which ostensibly contain more pertinent information about carbon flux, were the actual inputs into the clustering process. Even though the purchased maps are proprietary and cannot be distributed, the map products derived from them are not subject to this limitation (so long as the original map cannot be reverse-engineered back out).

We are also including monthly discrete counts of days with extreme events, *i.e.*, temperatures greater than 90°F or below 32°F. Days with measurable precipitation also fall into this category of counts of discrete events. Because of the limitations of the monthly data, we uniformly pro-rated these counts over the days of each month. For growing season periods beginning and ending on days other than month boundaries, we added these pro-rated partial sums over the appropriate number of days. In this way, total events were conserved and growing season counts plus non-growing season counts equaled annual counts. Including such extreme events (and the variance information that they convey) is important for carbon flux, even though these best-available data are temporally coarse.

1.6 Preparation of Input Layers for Multivariate Clustering

Before submitting them to the MGC procedure, all map layers were normalized by subtracting the mean from each observation, and dividing through by the standard deviation. This **normalization** is necessary, since absolute differences inherent in different units of measurement used for various map layers would otherwise unduly affect the clustering process. Only the variance information is used in the clustering process.

Consideration of this normalization process affected our treatment of a few map layers. For example, negative values of MODIS PSNNET indicate that maintenance respiration exceeds GPP, and this difference between positive values and negative values has particular biological significance. Since normalization of PSNNET would have destroyed the significance of this zero point where production exceeds respiration, we subtracted PSNNET from GPP to produce a respiration index (RI). Since both GPP and RI are always positive, and since PSNNET is related by subtraction, the inclusion of GPP and RI rather than PSNNET provides the same information and eliminates the possibility of destroying negative values of PSNNET.

1.7 Correlations Among Inputs

Before performing the statistical clustering to produce flux ecoregions, we examined the degree to which all of the input layers were cross-correlated with each other. Even if two maps show 95% correlation with one another, however, it may not be prudent to eliminate one of them. If the additional information carried in the uncorrelated 5% serves to distinguish or separate two regions which have important differences in carbon flux, and which would otherwise be confounded, then it is important to retain both layers in the analysis. Thus, the inclusion decision for input layers cannot be distilled to a simple correlation analysis. Ecological expertise must be used in considering which layers to include in the MGC analysis.

Thus, layers selected for inclusion were not be entirely uncorrelated with one another. The MODIS MOD17A3 GPP and NPP averaged annual products from 2001 and 2002 were found to be so highly correlated with GPP and RI seasonal layers that these two input layers were dropped. Some remaining input layer pairs showed expected high correlations greater than 0.9: cold days with cold sum, wet days growing with precipitation growing, solar growing with cold days (negatively), LAI non-growing with FPAR non-growing, GPP growing with FPAR growing and LAI growing, RI growing with LAI growing, and RI non-growing with GPP non-growing. The highest of these was 0.96, most were below 0.93.

Cross-correlations do not strongly affect the MGC results, but are best avoided if possible. **Principal Component Analysis (PCA)** could be used to produce axes which are orthogonal (uncorrelated) to each other for clustering. However, results from clustering within a PCA space will be in PCA units, and cannot be transformed back into the original units of measurement if PCA axes that explain even small amounts of variance have been dropped.

Most applications of PCA are designed to reduce dimensionality by eliminating all but the three or four axes which explain the majority of the variability in the data. Although accounting for the majority of variability, less than 100% of the variance is explained by any subset of the axes. This loss of explanatory power means that it is impossible to convert values from the reduced PCA space back into the original units.

We used a unique statistical technique to ensure not only that the axes forming the data space within which the MGC procedure operates are orthogonal, but also that we can untransform the clustered units back into original measurement units for each axis. We performed a PCA analysis on all input layers included in each flux ecoregionalization, but retained all of the PCA axes in the analysis. Thus, if there were 26 input layers, we retained all 26 PCA axes that resulted. All PCA axes produced were guaranteed to be orthogonal and uncorrelated, but, because 100% of the variance in the original data was retained, we could still convert these PCA scores back into units that are ecologically relevant.

1.8 Resolution of Printed Maps

This report is severely limited by the size of its printed pages. Small printouts cannot show the full resolution contained within these maps. Each map would have to be printed as a wall-sized poster to easily see the single square kilometer-sized cells that they contain. However, digital files for each of the 48 high-resolution maps in this report would each be 85 MB at full resolution. As a compromise between size and quality, we have printed maps with a resolution degraded to 1500m rather than the full 1000m resolution. It is difficult to see the difference between these in a 10 inch map printed on a standard 600 dpi printer.

This document has been released in PDF and in PostScript formats. The PDF version is much smaller and can be emailed easily, but the maps are resampled. If printed, they will appear to be much coarser than they actually are. The PostScript format is much larger, but will produce maps of 1500m resolution if printed. Versions of this document in both formats are available at http://geobabble.ornl.gov/flux-ecoregions. Figure 1.4 shows a closeup of Nevada at full 1km^2 resolution as an example. The aspect of individual mountain ranges can be seen as differences in total solar insolation during the non-growing season in this figure. This layer carries information about slope and aspect into the flux ecoregion process.

Also available at http://geobabble.ornl.gov/flux-ecoregions are a variety of supporting imagergy, animations, and GIS maps. Maps of the homogeneous flux ecoregions and the interpolated and integrated input layers can be viewed using the Pan/Zoom/Scroll tool on the website at http://geobabble.ornl.gov/cgi-bin/pzs?conf=FluxEcoregionMaps/FluxEcoregionMaps.conf. Animations of the synoptic interpolated time series input data derived from MODIS products are available in AVI, QuickTime, and MPEG formats. Digital GIS map layers of the resulting homogeneous flux ecoregions can be downloaded also.

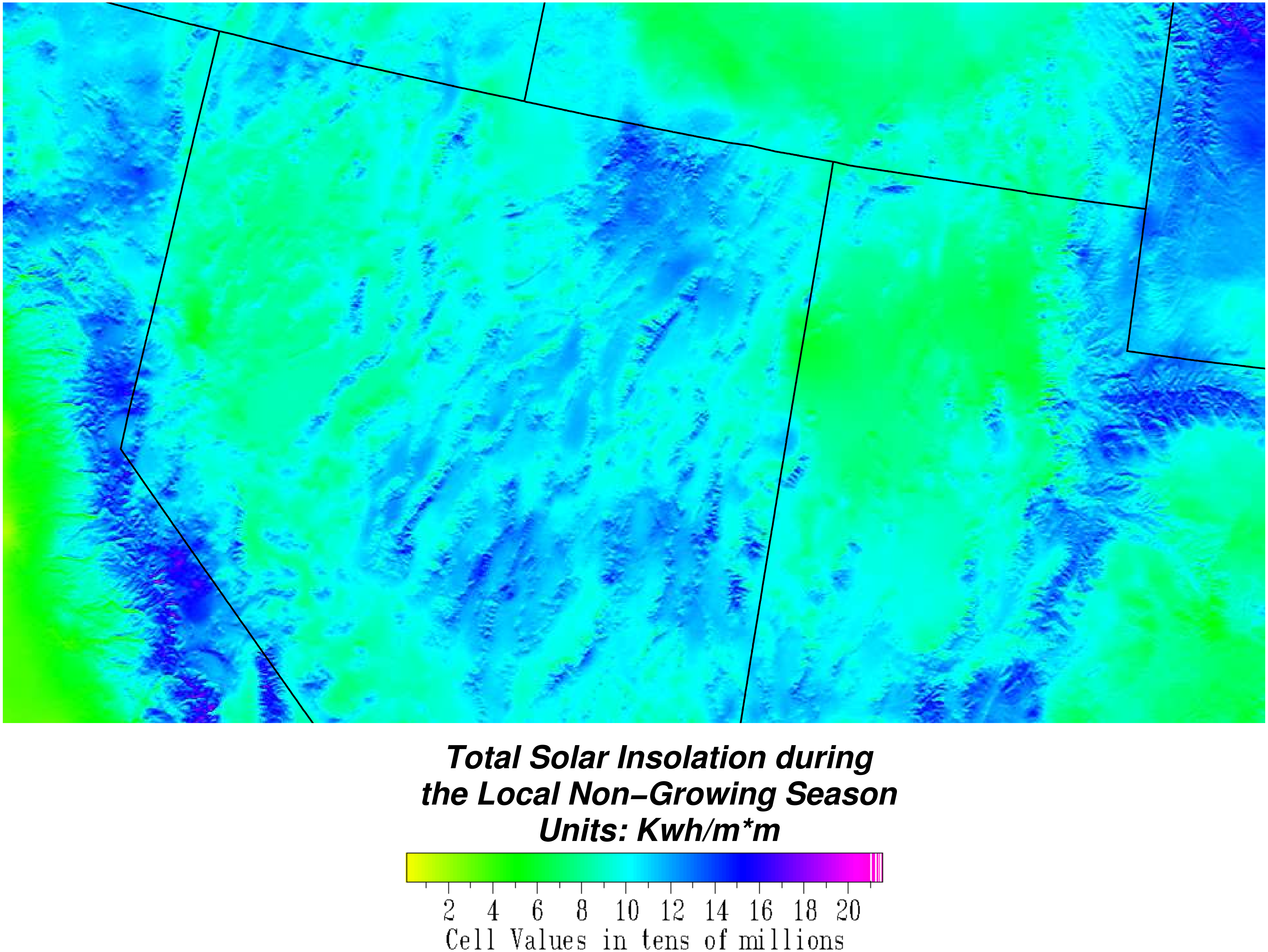

Figure 1.4: Total Solar Insolation During the Local Non-Growing Season in the Central Western United States