EcoTrends: past, present & future – an opportunity to promote synthetic analyses

Christine Laney, project coordinator
Jornada Basin LTER

Acknowledgements

Debra Peters, project leader, lead PI at Jornada Basin LTER & lead scientist at Jornada Experimental Range, USDA ARS

Ken Ramsey, information manager, JRN

Jin Yao, spatial analyst, JRN

Funding agencies: National Science Foundation (NSF), New Mexico State University (NMSU) and the USDA Agricultural Research Service (USDA-ARS) Jornada Experimental Range.

Collaborators: Long-Term Ecological Research (LTER) Network Office, USDA-ARS, USDA Forest Service (USDA-FS), National Center for Ecological Analysis and Synthesis (NCEAS), Department of Energy (DOE), US Geological Survey (USGS), and state institutions.

Scientists, information managers, support staff, and students from each site.

Editorial Committee: Debra Peters [JRN, ARS], Ariel Lugo [LUQ, USFS], Scott Collins [SEV], Charley Driscoll [HBR], Peter Groffman [HBR], Morgan Grove [BES, USFS], Alan Knapp [SGS, KNZ], Tim Kratz [NTL], Christine Laney [JRN], Mark Ohman [CCE], Bob Waide [LUQ, LNO]

Technical support: Ken Ramsey [JRN], James Brunt, Duane Costa, Inigo San Gil, Mark Servilla; Don Henshaw [AND], Wade Sheldon [GCE], Mark Schildhauer [NCEAS]

Background

- **Long-term ecological research sites**: 1902 – present, including LTER sites, USFS Experimental Forests, USDA Agricultural Research Service sites.

- These sites represent a **wide range of ecosystem types**
  - Forests to grasslands and shrublands
  - Freshwater lakes and streams
  - Near-coastal marine and estuaries
  - Urban areas
  - Arctic, Antarctic, and alpine

- **Kinds of available data**:
  - Climatic and demographic data since the 1800s
  - More recent quantitative assessments of plant, animal, and microbial populations and communities, hydrological and biogeochemical cycles, biodiversity, and disturbance regimes.

As the number and variety of long-term datasets grow, there is an acute need to make these data available for cross-site synthesis projects!

Challenges to cross-site analyses

1. **Few standardized raw datasets, few derived datasets and incomplete metadata**
2. **Limited data management tools** to track project information, source and derived datasets, metadata, and scripts.
3. **No website to support cross-site synthesis** that provides derived datasets that are easy to browse, explore, download and plot.
4. **Limited communication webs** between ecologists, information managers & computer scientists for improving access to clean data.

Goals

- **Create a platform for synthesis by making long-term data easily accessible**
- **Illustrate the utility of this platform in addressing within-site and network-level scientific questions.**

- **Wide-ranging audience!**

Products

- **Book** (Peters et al., 2010) to be published on trends in long-term data within and among sites, and examples that illustrate the value of long-term data in addressing important questions
- **Web site** containing derived data and metadata that can be easily explored, accessed, downloaded, and plotted for synthetic analyses.

Sites & Themes

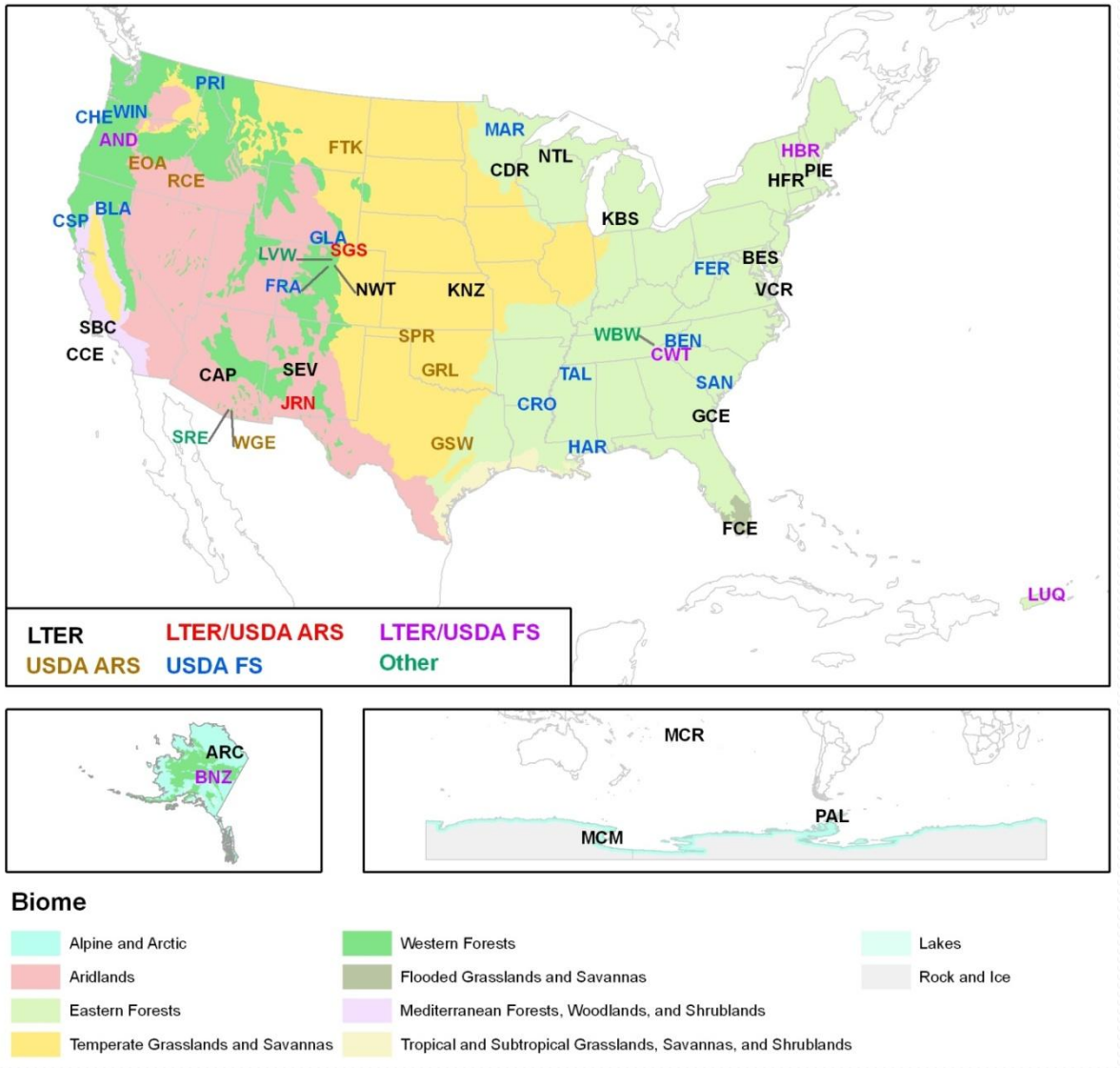
Biotic structure

Biogeochemistry

Disturbance

Human population & economy

Climate & physical variability



Early Years

- Requirements: kept low to retain inclusiveness
  - Datasets span at least 10 years (except for newer sites)
  - Each dataset should be clean enough to work with
  - Each dataset should be documented at a basic level (EML good, but not required – fairly new effort).
- 2004: early ideas and probing
  - Unknown what types of long-term data were available
  - Unknown how clean the data would be
  - Unknown how well-documented the data would be
- 2005: Attempted to obtain data via site websites.
  - Filtering out long-term data difficult
  - Some data that was known about was not online

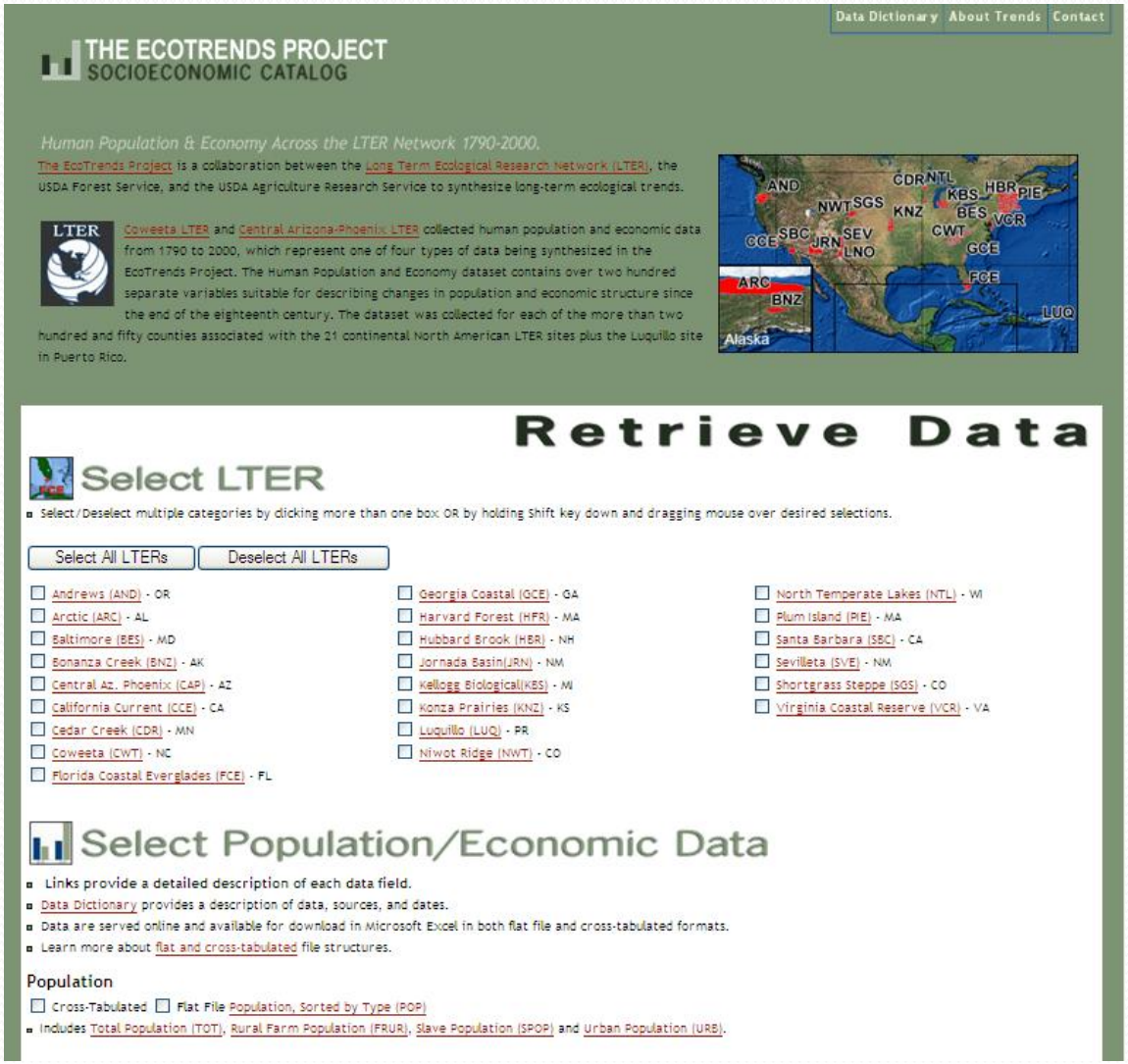
Data

- 50 sites contributed ~400 datasets (~1000 data files) to the project, of which ~90% come from the LTER community. These datasets were supplemented with data from the
  - National Atmospheric Deposition Program (NADP)
  - National Climatic Data Center (NCDC)
  - Climate/Hydrology Database Projects (CLIMDB)
  - National Oceanic and Atmospheric Administration (NOAA)
  - Clean Air Status and Trends Network (CASTNET)
  - United States Geological Survey (USGS)
  - The EcoTrends Project Socioeconomic Catalog (separately funded effort by T. Gragson, C. Boone, M. Grove; data from US Census Bureau)

Socioeconomic catalog

http://coweeta.uga.edu/trends/catalog_trends_base2.php

- 2006-2007
- 23 of the 26 LTER sites
- 800 counties (including complete air sheds of some sites)
- ~30 variables, from total population to # of service establishments.



Data

Theme	Variables
Biotic	animal abundance; animal and plant biomass, cover, density, species richness; primary productivity; chlorophyll A, etc.
Disturbance	fires, storms
Human	total population, population density, percent urban population, percent of total population employed in farming, commerce, manufacturing and service
Biogeochemistry	ammonium, calcium, chloride, nitrate and sulfate (and others) in precip. and surface water
Climate/physical variability	air temp., precipitation, PDSI, wind speed, mean sea level, soil temp., water temp., secchi disk depth, streamflow, ice duration, solar radiation

Working with the data: 2005-2009

Submitted data and metadata formats

- Spreadsheets: every one has a unique layout
- ASCII files: includes tab, comma, pipe-delimited and fixed-width formats
- Word processing documents
- Adobe .pdf files
- Published papers
- Statistical software files (e.g., Sigmaplot, SAS)
- Relational database files (e.g.,.dbf, mdb)
- Ecological Metadata Language (EML)

Working with the data

First lesson: data + metadata does not necessarily equal 'good' data!

Data quality

- Column headers: poorly documented
- Quality flagging: un- or poorly defined
- Data quality checking: typos, incorrect values, lack of standardized factor names, blank lines, duplicate rows, etc.
- Metadata embedded in data document

Metadata quality

- Documentation of data tables: Non-existent, incomplete, or incorrect documentation prevents automation of data derivation.
- Documentation of methodology: Non-existent, incomplete, or incorrect documentation of data collection, quality-checking, and other preprocessing procedures can cause inappropriate aggregation and interpretation of the data.

Intra-site data challenges

Example: A productivity/physical conditions dataset with one file per year: note different formats and column headers (which reflect a dynamically evolving study)

1983

7	Lake	Code	DATE	DEPTH	temp C	pH	cond	O-2	alk	LIGHT	SECCHI	1 prod
8				meters			umhos/cm	mg O-2/l	meq/l	u Einst/	meters	
9										cm^2-sec		
10	Toolik	100	13-Jul-83	0	11.7	7.42	35	11.25				21.019
11	Toolik	100	13-Jul-83	1	11.9	7.435	38	11.1				11.39
12	Toolik	100	13-Jul-83	3	11.9	7.46	38	11				5.166
13	Toolik	100	13-Jul-83	5	11.8	7.49	38	11				0.406
14	Toolik	100	13-Jul-83	8	6	7.365	35	12.3				1.519
15	Toolik	100	13-Jul-83	12	4	7.26	43	10.9				0.294
16	Toolik	100	13-Jul-83	16	3	7.155	45	10.1				0.198

2004

1	site	Lake	Date	Depth	Temp	pH	SpCond	DO	chlor	Light (sub)	Light (deck)	secchi	Birgean	Extinction or absorption	vertical ext coef (nt)	Notes
2	code		MMDDYY	meters	degC	units	uS/cm	mg/l	volts	u einst	u einst	meters	% absorption	co efficient		
3																
4	100	Toolik Main	18-Jun-04	0.09	5.53	6.81	25	10.38	0.089	1248	1280.8	4.95				
5	100	Toolik Main	18-Jun-04	0.5	5.52	6.81	25.1	10.24	0.088	747		4.95				
6	100	Toolik Main	18-Jun-04	1.01	5.54	6.82	24.9	10.2	0.091	114		4.95				
7	100	Toolik Main	18-Jun-04	1.48	5.51	6.82	25	10.12	0.097	357		4.95				
8	100	Toolik Main	18-Jun-04	1.99	5.52	6.81	25.1	10.12	0.107	163		4.95				
9	100	Toolik Main	18-Jun-04	2.49	5.51	6.82	25.1	10.08	0.129	95		4.95				
10	100	Toolik Main	18-Jun-04	3	5.52	6.82	25	10.1	0.14	82		4.95				

Inter-site data challenges

Site A

Exp	Year	OldField	Plot	Transect	YearAb	Sr	TotBio	ShanWinr	Evenness	EtoH	AdjSR	Moisture	NO3Soil	Acerrubr	Achimill	Agroalba	Agrorepe	Agroscat	Ambrart
54	1988	24	1	G	1968	2	41.4333	0.6195	0.8938	1.8581	2	4.58	0.2	0	0	0	12.867	0	0
54	1988	24	2	R	1968	7	58.4	0.7735	0.3975	2.1673	5	4.6	0	0	0	0	0	0	0
54	1988	24	3	W	1968	6	96.3667	0.6023	0.3362	1.8264	3	7.08	0	0	0	0	18.7	0	0
54	1988	24	4	Y	1968	9	158.033	1.145	0.5211	3.1424	6	4.85	0	0	0	0	0	0	0.167
54	1988	26	1	G	195	6	168.433	1.0956	0.6115	2.9011	5	6.33	0	0	0	0	4.067	0	0
54	1988	26	2	R	195														
54	1988	26	3	W	195														
54	1988	26	4	Y	195														
54	1988	35	1	G	194														

Site B

year	site	trt	web	plot	quad	species	wim
1999	C	U	1	E	1	CHSE7	0.00
1999	C	U	1	E	1	CHSES	0.00
1999	C	U	1	E	1	DAPU7	0.12
1999	C	U	1	E	1	GUSA2	0.79
1999	C	U	1	E	1	LEFE	0.37
1999	C	U	1	E	1	THAC	0.00
1999	C	U	1	E	2	DAPU7	0.01
1999	C	U	1	E	2	GUSA2	0.01
1999	C	U	1	E	2	LATR2	235.12

	A	B
Variables	Aboveground ***annual*** biomass, with many contextual variables (e.g., soil moisture & N content) and summary variables (e.g., species richness). EcoTrends does not include some of these variables.	Aboveground ***seasonal*** net primary productivity. EcoTrends aggregates to annual (does not include seasonal data)
Experimental design	Data collected in old fields with several transects each; each transect has several sampling plots	Data collected under a complex web design in different vegetation zones with different treatments.
Species codes	Uses codes specific to site studies	Most codes are from USDA PLANTS, except for codes for litter and unidentified species.
Format	Tab-delimited (shown here as .xls)	Space-delimited

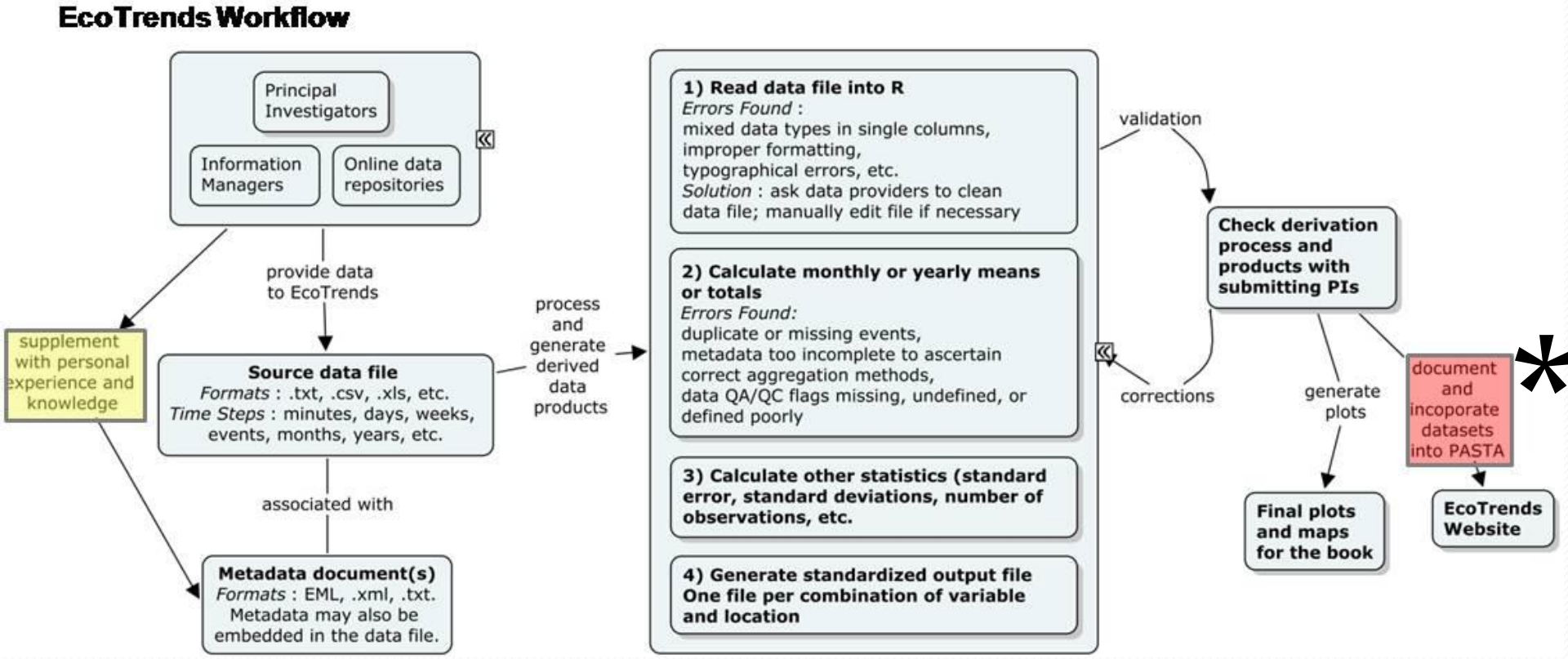
Final derived data product

One derived format allows for integration of 1000's of datasets

site	station	year	month	observation
JRN	JRN-Creosote	1990	NA	94.5
JRN	JRN-Creosote	1991	NA	82.9
JRN	JRN-Creosote	1992	NA	112.5
JRN	JRN-Creosote	1993	NA	85.2
JRN	JRN-Creosote	1994	NA	55
JRN	JRN-Creosote	1995	NA	73.4
JRN	JRN-Creosote	1996	NA	109.7
JRN	JRN-Creosote	1997	NA	65.5
JRN	JRN-Creosote	1998	NA	62.8
JRN	JRN-Creosote	1999	NA	93
JRN	JRN-Creosote	2000	NA	42
JRN	JRN-Creosote	2001	NA	71
JRN	JRN-Creosote	2002	NA	65.3
JRN	JRN-Creosote	2003	NA	41.2
JRN	JRN-Creosote	2004	NA	74.6
JRN	JRN-Creosote	2005	NA	92.5
JRN	JRN-Creosote	2006	NA	111.8
JRN	JRN-Creosote	2007	NA	110.9
JRN	JRN-Creosote	2008	NA	147.8

… other summary statistics (not displayed on web)

Dynamic, engaged workflow



Products

Book

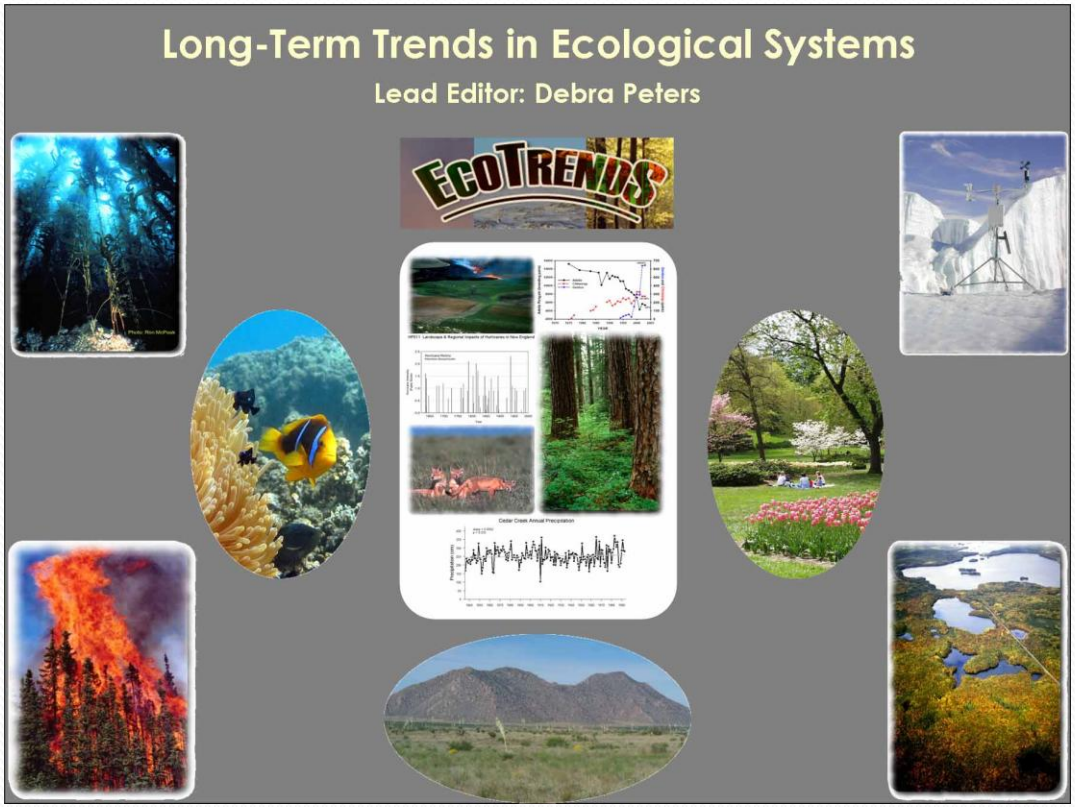
EcoTrends Website

Socioeconomic Catalog Website

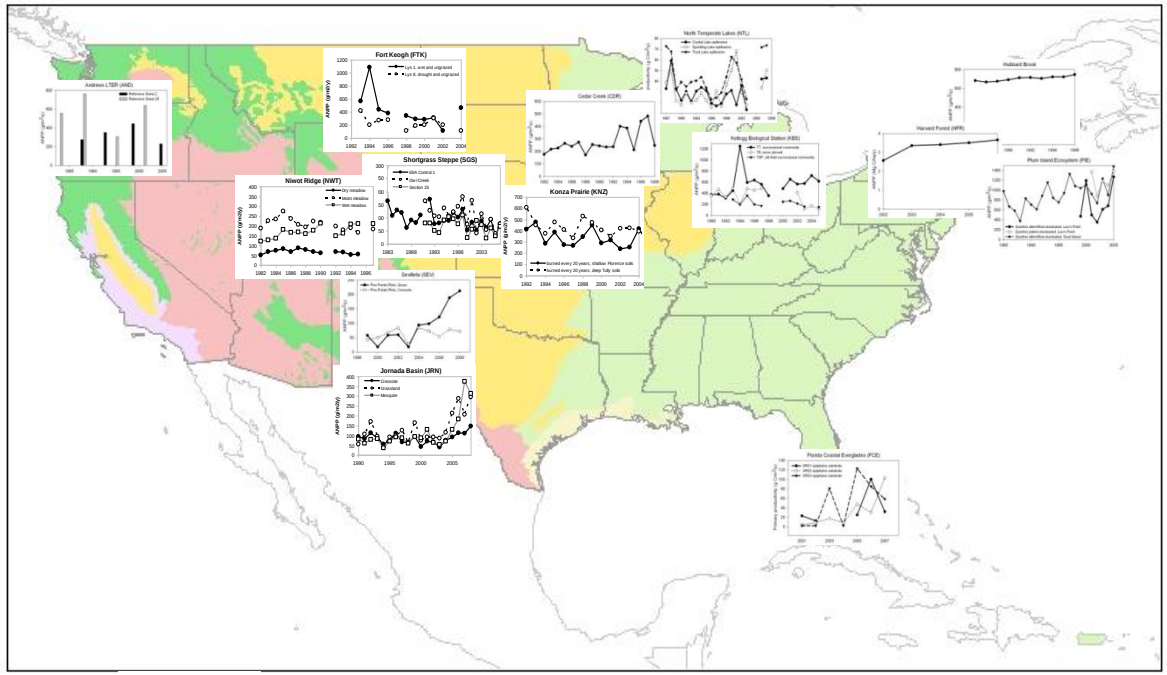
P²ERLS

Implementing the EcoTrends Vision

Book in final stages - look for it in 2010



Biotic structure



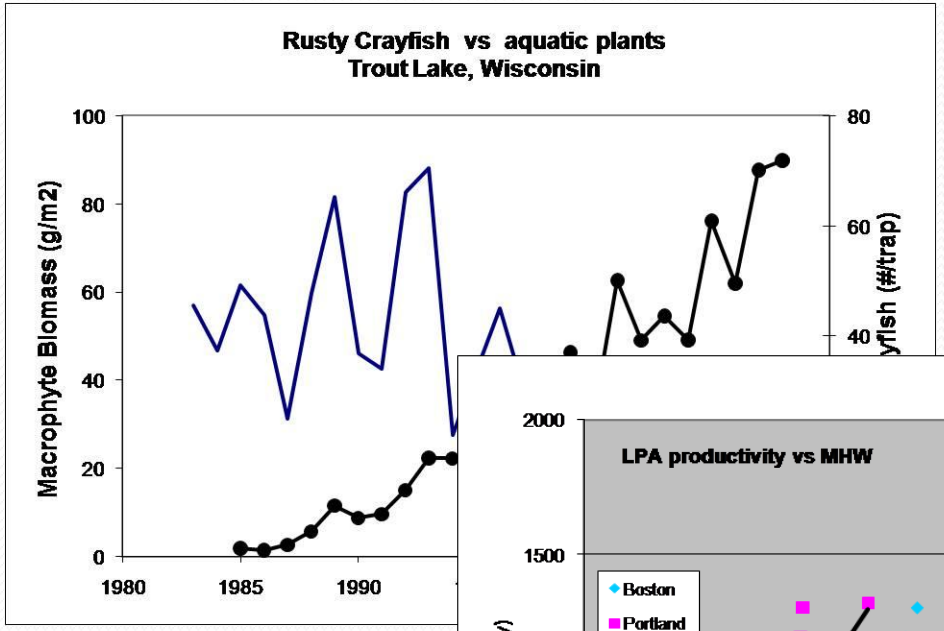
Variables

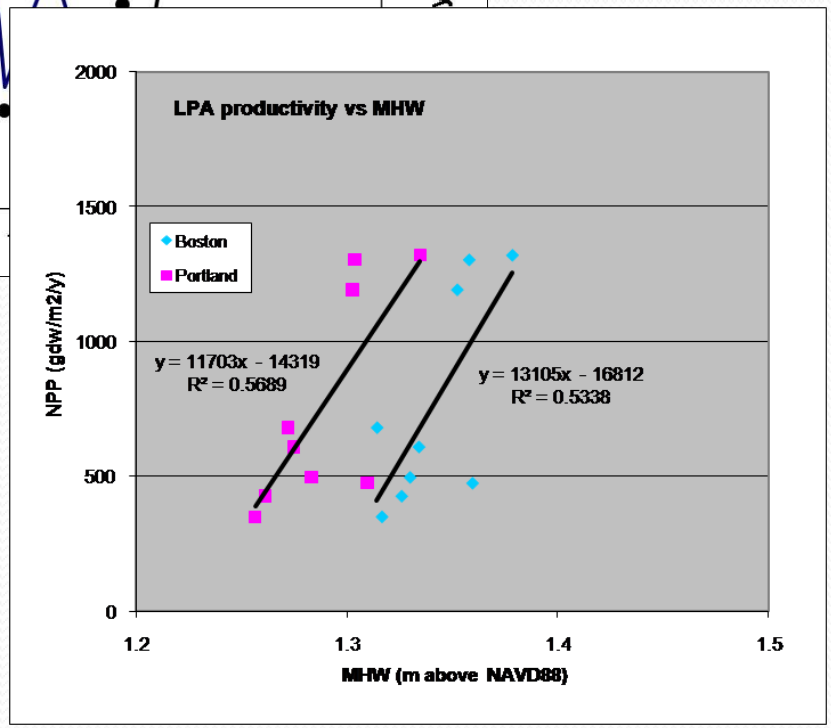
- animal abundance
- animal and plant biomass, cover and density
- species richness
- primary productivity
- chlorophyll A, etc.

Disturbance

Variables

Biotic response to
- fires
- storms
- landslides
- human
- population/landuse
- invasive species



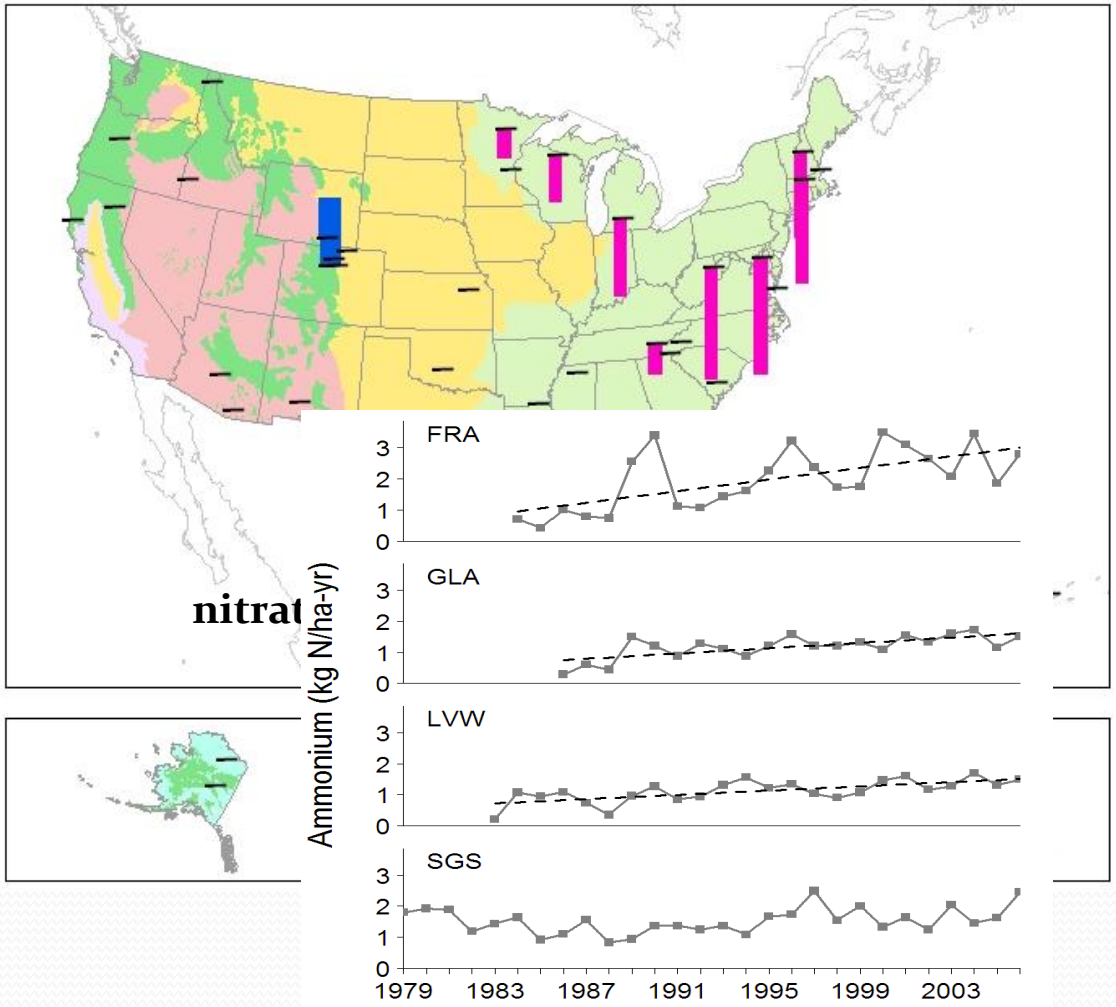


Biogeochemistry chapter

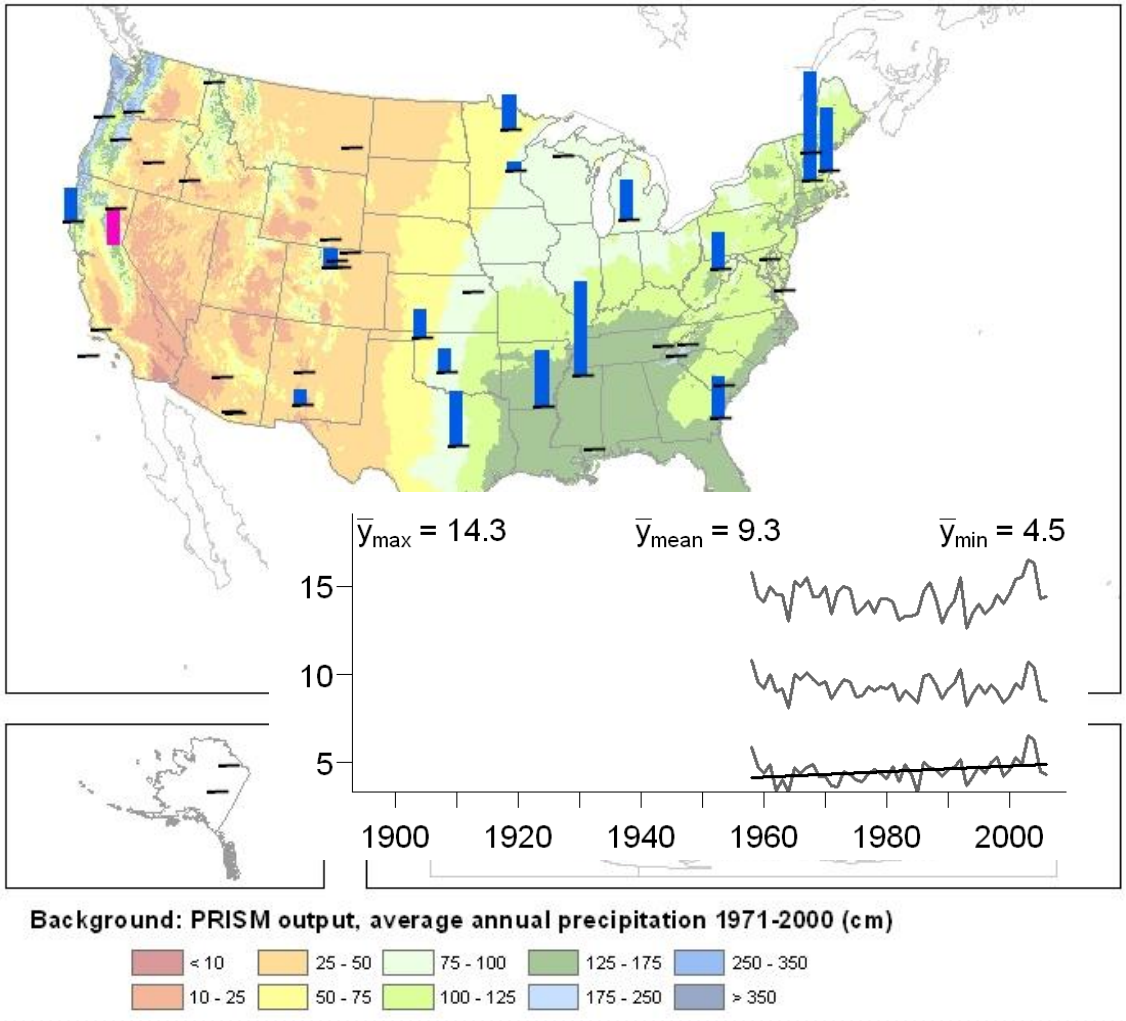
Variables

Calcium
Chloride
Ammonium
Nitrate
Sulfate

Precipitation
Wet deposition
Stream flux
Surface water concentrations



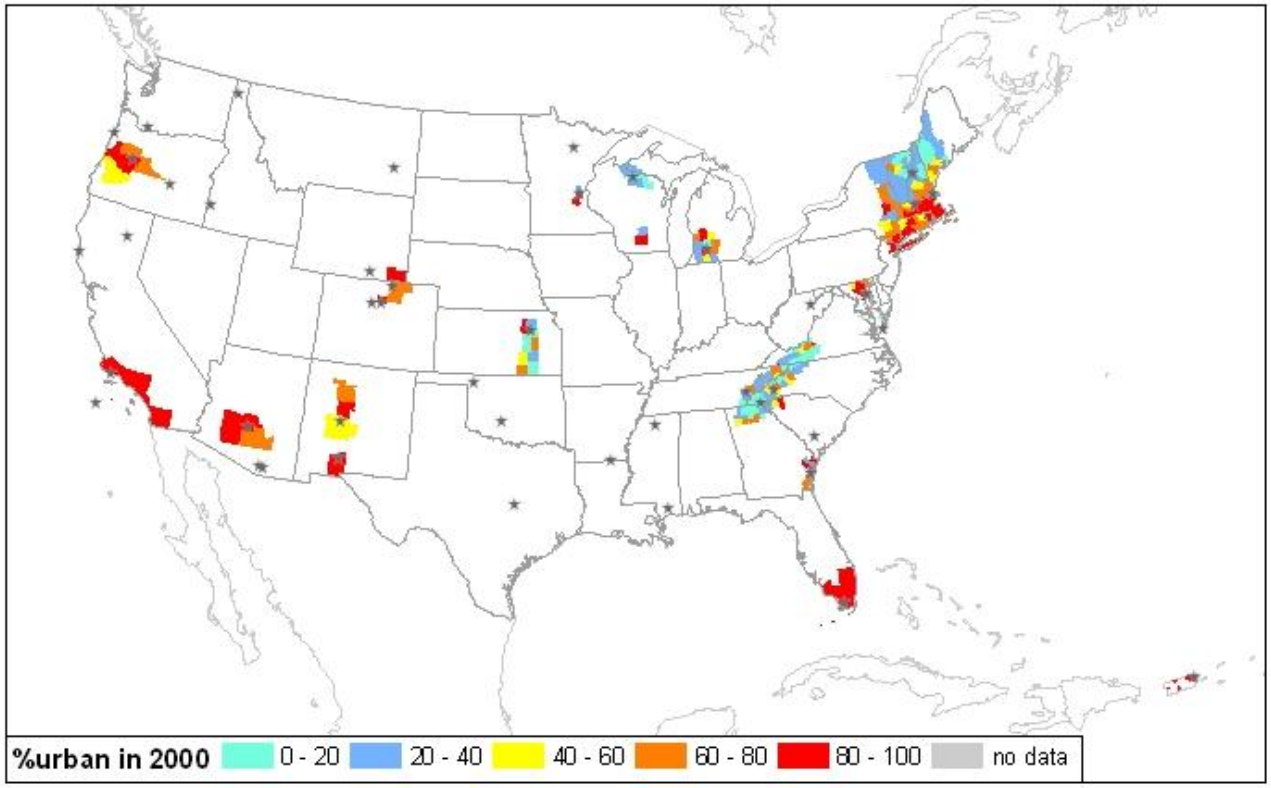
Climate chapter

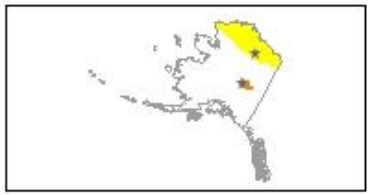


Variables

Air temperature
Precipitation
Palmer drought severity index
Wind speed
Soil temperature
Ice duration
Etc…

Human Population & Economy

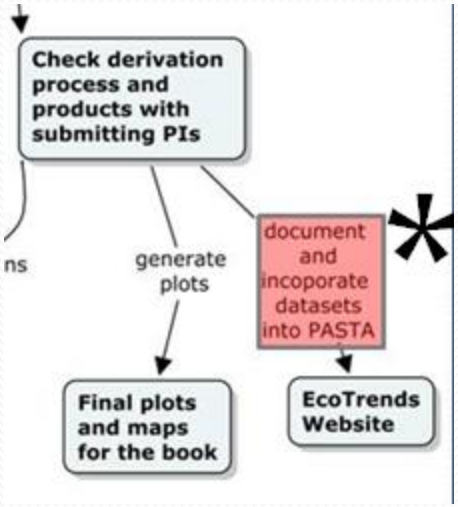




Variables

Total population
County area
Population density
Percent urban
% employed in
farming,
commercial,
manufacturing, &
service

Website – www.ecotrends.info

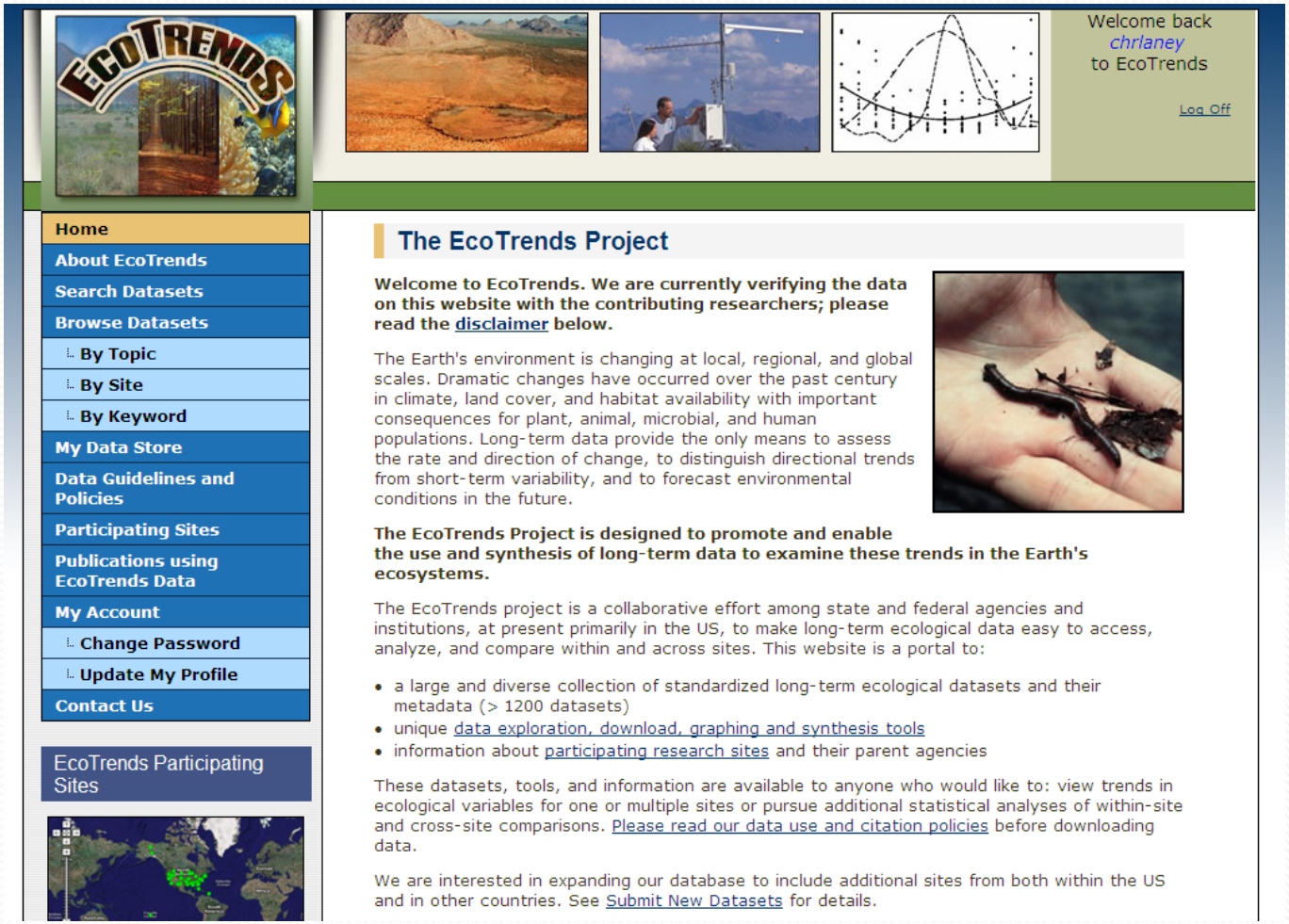


Servilla et al. 2008

- Datasets are read into a indexed database
- Information from the EcoTrends database table describing
  - the origin of the datasets
  - links between files
  - information about the sites and stations
  - Etc.

  are utilized to create EML documents describing the derived data products
- The combined system provides data to the website

Website – www.ecotrends.info



And P2ERLS (www.p2erls.net)

Impact & Lessons Learned

Undergraduate educator workshops with ESA & NEON

Science working groups

Informatics/Technology working groups

ESA & NEON undergraduate educator workshops

- **Using Continental-scale Data to Teach Undergraduate Ecology**
- 2 workshops (DC in Oct 2008, Albuquerque in Aug 2009)
- NEON – developing a conceptual framework to guide decisions about how make their data accessible to educators when it become available.
- EcoTrends provides a working example
- Educators provided feedback to EcoTrends

Lessons learned from educators

- **Need another interface for students and teachers**
  - Lots of guidance
  - Limited datasets
  - Links to curriculum centered around particular datasets
- **Important to be able to zoom in and out**, both **temporally** (effects of short vs. long datasets) and **spatially** (environmental context).
- **Main topics of interest**:
  - relationship between human population density and nutrient/pollutant flux
  - Effects of climate change on ecosystem communities
- Important for students to learn about cleaning data, reading metadata, plotting and analyzing data.
- Certain set of skills needed to be able to interact with EcoTrends: computer literacy, science literacy, statistics, etc.
- **More data, please!** (e.g., CO_2 levels, evapotranspiration, etc.)

Science working groups

- Series of meetings in April/May
  - Disturbance: Peters
  - State change: Bestelmeyer
  - Biogeochemistry: Driscoll & Groffman
  - NPP & species richness: Collins
  - Animal populations: Willig
  - NOVEMBER Socioeconomic: Grove
- Decided on a science question or conceptual model to explore
- Explored EcoTrends data availability and accessibility
- **Most groups were able to quickly access and assess data, and use their findings to guide the direction of their discussions.**

Science working groups

- Some data still needed to be checked by Pis
- Some data were missing
  - Data were unavailable online or not submitted to EcoTrends
  - Data could not be incorporated into EcoTrends using the current data model
- Interface needs some work: not always intuitive.
- Identified a set of future challenges that should be met to meet the original vision of the project.

Problem #1: Making all long-term data available

- **Event-driven data** that should not be plotted over calendar time, including LIDET and post-treatment data
- **Disturbance data**
  - Incorporating better information about disturbance drivers/events
  - Linking driver and response datasets
- **Species data**
  - Available by species and community
  - Datasets linked by common source
  - Species identifiers standardized so that they can be compared across sites.
- **Experimental data**
  - mostly 'reference' datasets
  - System not best suited to incorporate information about treatments or varied environmental context.

Problem #1: Making all long-term data available

Possible future solutions:

- Restructure and expand the EcoTrends database to
  - Incorporate event-driven data
  - link species data to standardized, national databases (e.g., ITIS)
  - include more experimental data
- Utilize the database structure to show relationships between derived datasets on the website
- Link disturbance driver and response datasets with each other
- Allow users to view all species at once, or view one species at a time with links back to the larger dataset.

Problem #2: Plotting any X vs. Y and downloading multiple datasets together

- Data can only be plotted with time on the X axis.
- Users should also be able to dynamically plot any two variables against each other (eg., PPT and ANPP)
- Users should be able to download multiple datasets together in one file.

Possible future solutions:

- Redesign plotting interface to allow multiple-variable plotting
- Create tools that will allow users to download multiple datasets with the appropriate metadata in a single package.

Problem #3: Integrating temporal and spatial data

- Multiple locations are often sampled within a site, and these data are abstracted from their spatial context (topographic heterogeneity, distance between sites, aspect, elevation, etc.).
- Temporal data and their spatial context need to be presented together.

Possible future solutions:

- Expand the data model, database, and website to allow users to select data from a map rather than a list when desired
- Return maps with available datasets highlighted on a map after a query.
- Explore other mapping solutions (e.g., Arc Server , Arc Explorer). Use existing web services or create new ones to include more contextual data.

Problem #4: Coordinating between many sites & people

- Reporting errors back to the sites, and networking corrections between the site and EcoTrends databases
  - Help sites detect problems
  - Support future cross-site analyses
  - Support other LTER network systems like PASTA
- Networking with site IMs
  - Track future changes to data and metadata contents and their locations on websites
  - Keep EcoTrends file and database systems up-to-date.
- Coordinate & process scientific and information management ideas from emails, calls, and personal meetings.

Possible future solutions:

- A wiki site containing information about the sites
- Easier ways to check data and report errors
- Online data submission page that allows sites to directly submit derived data and metadata
- Social networking tools to communicate news on a frequent basis.

Science working groups

- **Disturbance: Peters**
  Working Group Session 5, Wed, 09/16/2009 - 10:00am - 12:00pm
  Reusch Auditorium Billhiemer
- **Biogeochemistry: Groffman & Driscoll**
  Working Group Session 6, Wed, 09/16/2009 - 1:30pm - 3:30pm
  Longs Peak Keyhole
- **NPP & species richness: Collins**
  Working Group Session 4, Tue, 09/15/2009 - 3:00pm - 6:00pm
  Longs Peak Granite Pass
- **State change: Bestelmeyer**
  Working Group Session 4, Tue, 09/15/2009 - 3:00pm - 6:00pm
  Reusch Auditorium Hobbs
- **Animal populations: Willig**
  Working Group Session 2, Mon, 09/14/2009 - 4:00pm - 6:00pm
  Longs Peak Keyhole

Informatics/Technical EcoTrends working groups

- **Integrating spatial and temporal data across the LTER network (Laney & Yao; Valentine & Vande Castle)**

  Working Group Session 3, Tue, 09/15/2009 - 10:00am - 12:00pm

  Longs Peak Chasm Lake

- **Building an LTER Taxonomic Database to Support Synthetic Research (Laney & Ramsey; Patterson, Zeman, O'Brien, Gastil-Buhl, Sheldon)**

  Working Group Session 6, Wed, 09/16/2009 - 1:30pm - 3:30pm

  East Side Deer Ridge Fireside

Breakout!

Work with a group (3-5 people) to address one problem (20 minutes)

Report back to working group (5 minutes each)

Round-table discussion (20 minutes)

- Making all long-term data available
- Plotting any X vs. Y and downloading multiple datasets together
- Integrating temporal and spatial data
- Coordinating between many sites & people

The resulting report will inform future EcoTrends development to meet your needs.