Overview of data analysis and tools used for high frequency data analysis

2nd MANTEL Training School, Neuglobsow, Germany, February 2018

Biel Obrador

University of Barcelona

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 722518.

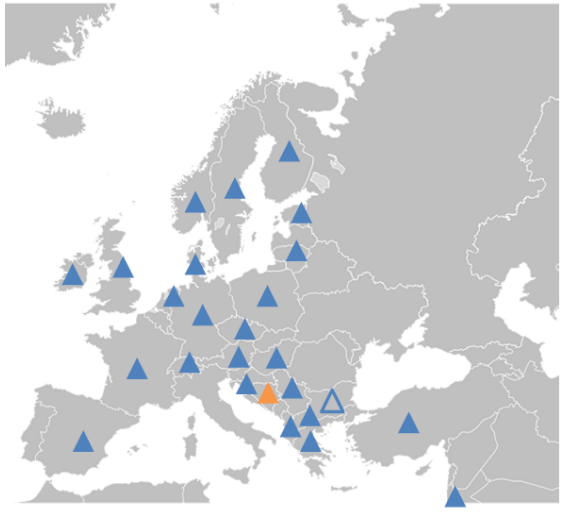
The NETLAKE COST Action

NETLAKE

cost EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

NETworking LAKe observatories in Europe

To build **a network of sites and individuals** that will support the development and deployment of sensor-based systems in lakes and reservoirs and promote the use of these systems to address both **current and future** water quality issues.



2012-2016

26 COST counties

4 non-COST
NZ, USA,
Australia, Albania.
EU: JRC, Ispra

NETLAKE WG2: DATA ANALYSIS AND MODELLING

TASK 1: collate existing high-frequency data analysis and modelling tools

NETLAKE WG2: DATA ANALYSIS AND MODELLING

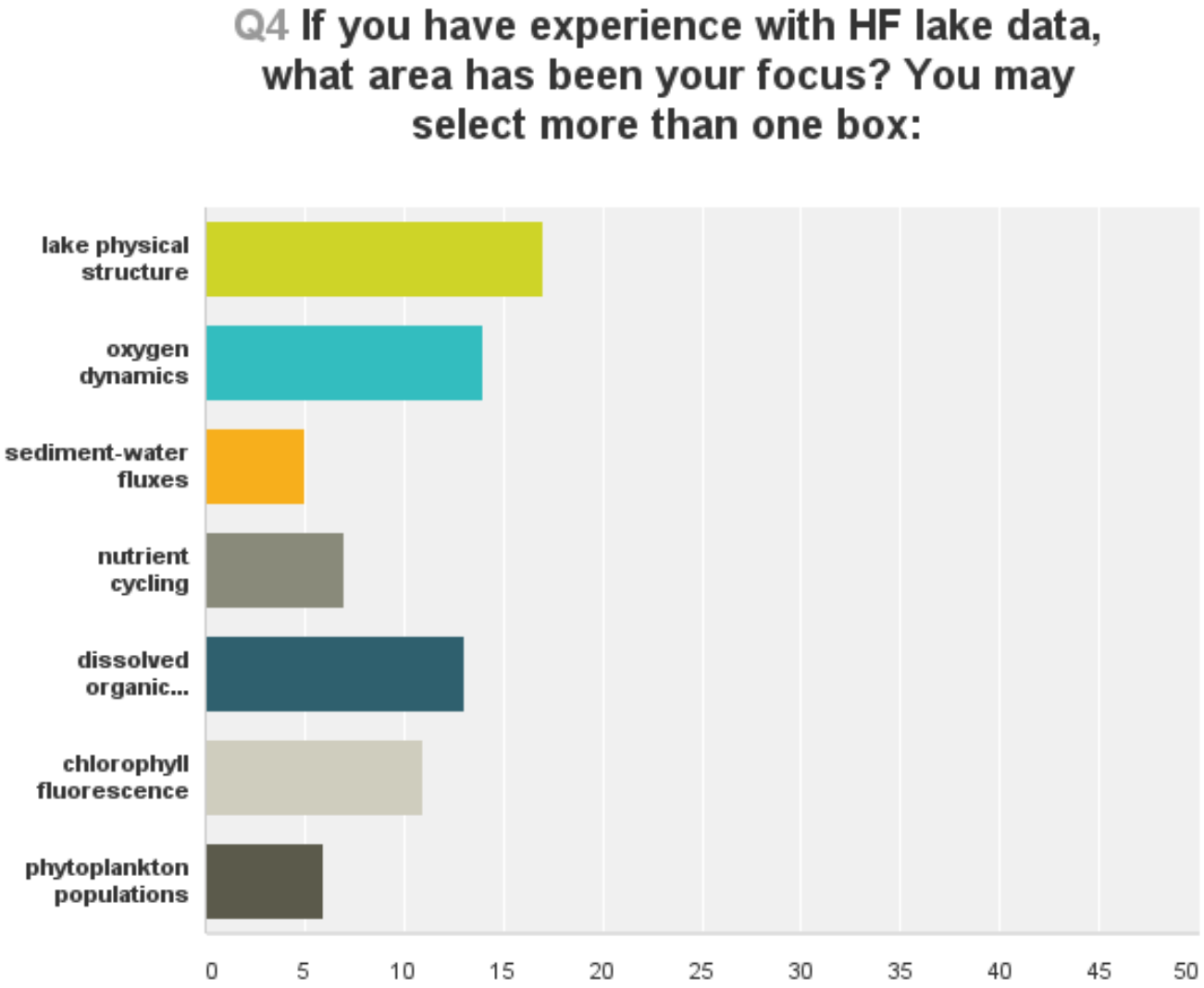
Survey within the NETLAKE community

- Aim to identify which tools were used by the experienced users of HF data
- Preliminary list of 18 tools identified (n=33)

Mugla, 2015

Girona, 2014

NETLAKE WG2: DATA ANALYSIS AND MODELLING



NETLAKE WG2: DATA ANALYSIS AND MODELLING

Tool type	Tool / model name	Program used	links	Reports, manuals or papers	Code availability
QA/QC	-	B3			Commercial software
	Handling big data sets	STATISTICA v7			Commercial software
		Data Mining Server	http://dms1.irb.hr/		?

@Mikkel Andersen

NETLAKE WG2: DATA ANALYSIS AND MODELLING

Tool type	Tool / model name	Program used	links	Reports, manuals or papers	Code availability
QA/QC	-	B3			Commercial software
	Handling big data sets	STATISTICA v7			Commercial software
		Data Mining Server	http://dms1.irb.hr/		?
Visualisation	None	JMP			Commercial software
	None	SAS			Commercial software
	None	Origin			Commercial software
	None	Sigmaplot			Commercial software
		Excel			Commercial software
		Graph Pad			Commercial software

@Mikkel Andersen

NETLAKE WG2: DATA ANALYSIS AND MODELLING

Tool type	Tool / model name	Program used	links	Reports, manuals or papers	Code availability
QA/QC	-	B3			Commercial software
	Handling big data sets	STATISTICA v7			Commercial software
		Data Mining Server	http://dms1.irb.hr/		?
Visualisation	None	JMP			Commercial software
	None	SAS			Commercial software
	None	Origin			Commercial software
	None	Sigmaplot			Commercial software
		Excel			Commercial software
		Graph Pad			Commercial software
Statistical tools A) Time series -Filtering -smoothing	Kalman	SAS, R, Matlab, STATISTICA		1, 2	Commercial/open access
	Wavelet	R, Matlab		3	Commercial/open access

@Mikkel Andersen

NETLAKE WG2: DATA ANALYSIS AND MODELLING

Tool type	Tool / model name	Program used	links	Reports, manuals or papers	Code availability
Lake calculating and modellig tools		myLake v1.12			Open Access
		Lake Analyser	http://lakeanalyzer.gleon.org	4	Open Access
		DYRESM CAEDYM			Commercial software
		pTOLOMEE			Commercial software
		Matlab			Commercial software
		PCLake		5, 6	Open Access?
		GLM FABM			Open access
		Special design programs			

@Mikkel Andersen

NETLAKE toolbox

1. Survey within the NETLAKE community
2. Definition of contents and topics to be included, identification of volunteer experts

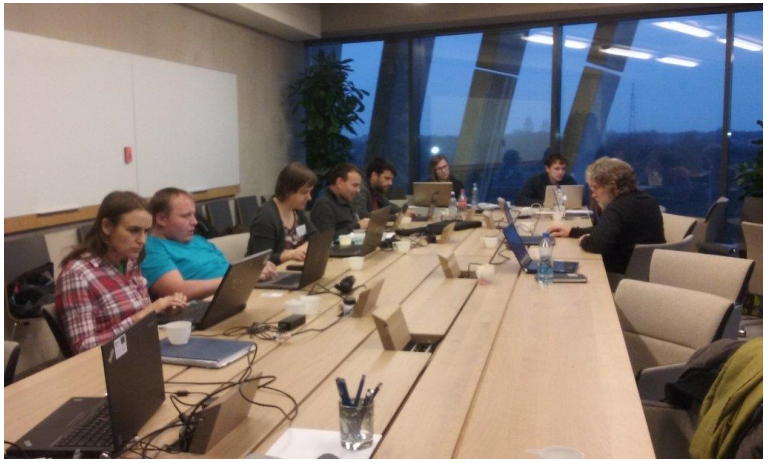
Uppsala, 2014

NETLAKE toolbox

1. Survey within the NETLAKE community
2. Definition of contents and topics and topics to be included, identification of volunteer experts
3. Writing
4. Reviewing (at least 3 reviews)

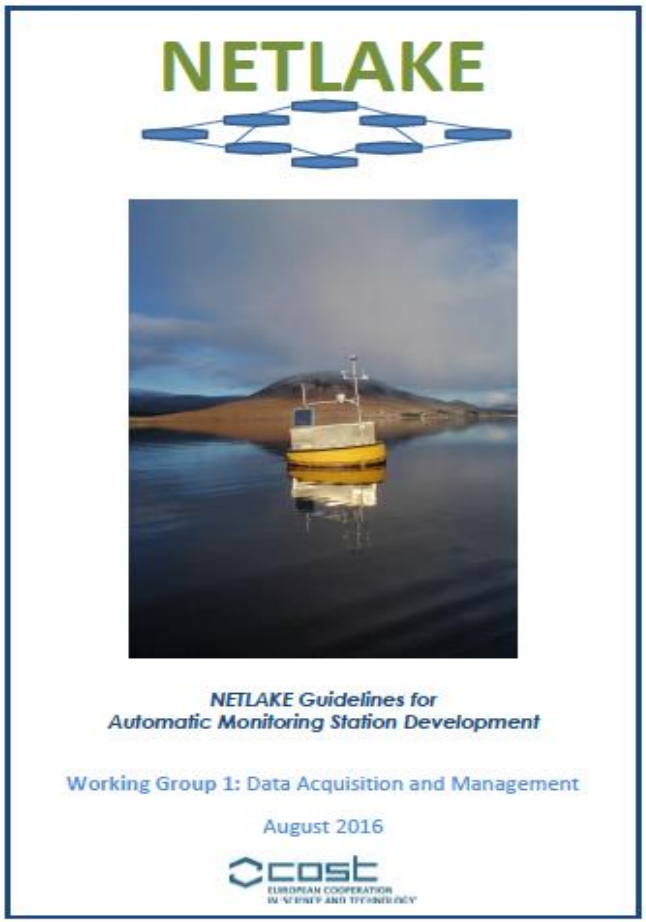
Evian, 2015

Riga, 2016

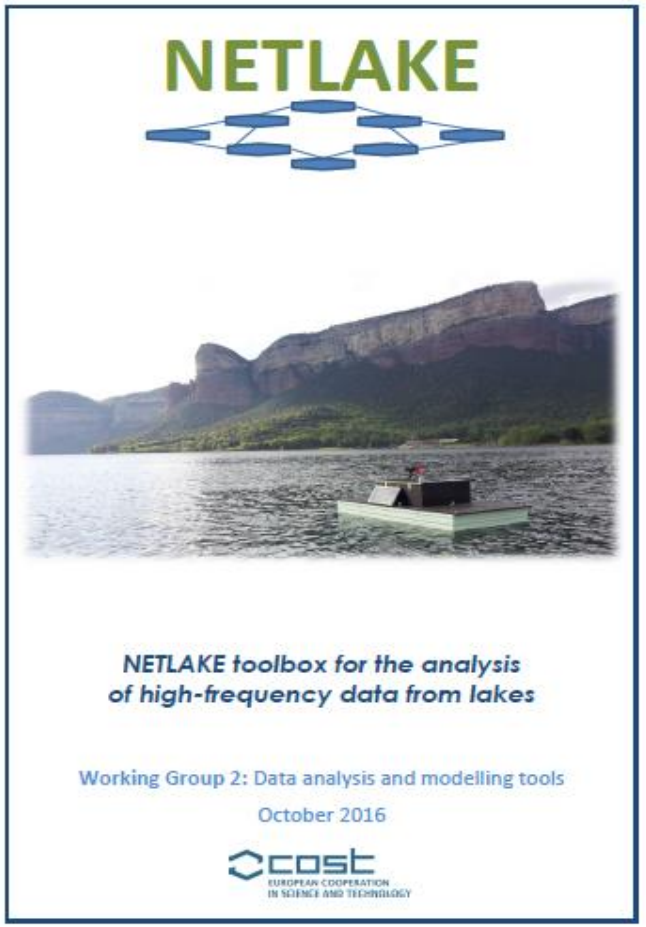


NETLAKE toolbox

www.dkit.ie/netlake



Laas et al. 2016



Obrador et al. 2016

The NETLAKE toolbox is

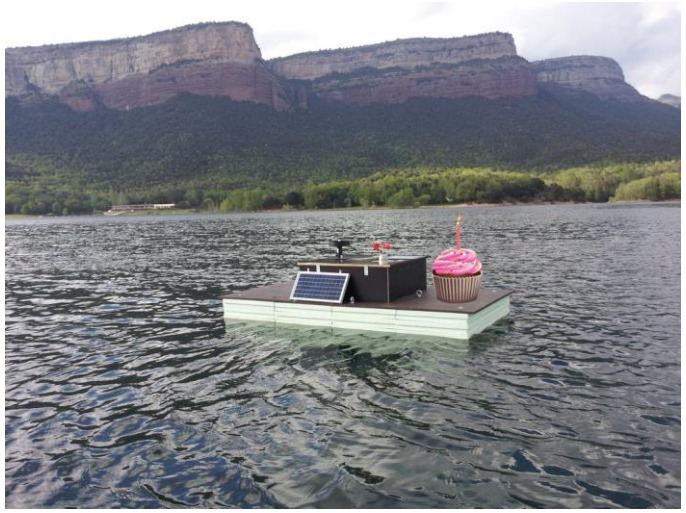
a collection of **short factsheets** on different **available tools** used in the processing, analysis and modelling of **high-frequency** lake data

is NOT

a review paper, nor a synthetic textbook

The primary target **audience** is

researchers and managers familiar with high-frequency lake monitoring data but **without previous experience** in each particular topic



The **function** of the factsheets is

to give unexperienced users an **introduction** to the **fundamentals** of a given technique and its **application**, taking advantage of experience of **specialized users** through

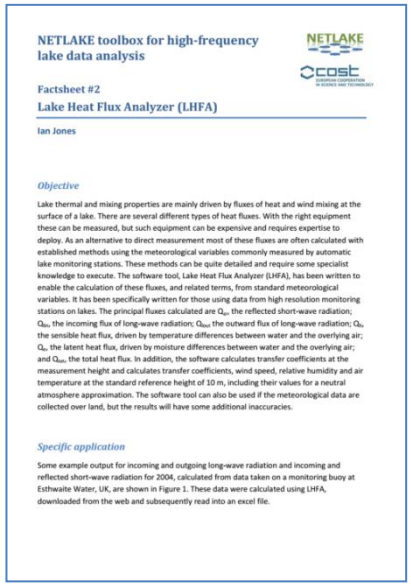
- listed pitfalls and tips,
- direct links to available **codes** and
- links to relevant scientific **literature**

NETLAKE toolbox

Each factsheet contains information on

- ✓ Objective and specific application
- ✓ Required knowledge, skills and resources
- ✓ Type of data and computational requirements
- ✓ Basic procedures
- ✓ Pitfalls and tips

- ✓ Contact details
- ✓ Reference literature
- ✓ Link to codes when available



NETLAKE toolbox for high-frequency lake data analysis

NETLAKE

cost

Factsheet #2

Lake Heat Flux Analyzer (LHFA)

Ian Jones

Objective

Lake thermal and mixing properties are mainly driven by fluxes of heat and wind mixing at the surface of a lake. There are several different types of heat fluxes. With the right equipment these can be measured, but such equipment can be expensive and requires expertise to deploy. As an alternative to direct measurement most of these fluxes are often calculated with established methods using the meteorological variables commonly measured by automatic lake monitoring stations. These methods can be quite detailed and require some specialist knowledge to execute. The software tool, Lake Heat Flux Analyzer (LHFA), has been written to enable the calculation of these fluxes, and related terms, from standard meteorological variables. It has been specifically written for those using data from high resolution monitoring stations on lakes. The principal fluxes calculated are Q_{sr}, the reflected short-wave radiation; Q_{lin}, the incoming flux of long-wave radiation; Q_{lout}, the outward flux of long-wave radiation; Q_h, the sensible heat flux, driven by temperature differences between water and the overlying air; Q_e, the latent heat flux, driven by moisture differences between water and the overlying air; and Q_{tot}, the total heat flux. In addition, the software calculates transfer coefficients at the measurement height and calculates transfer coefficients, wind speed, relative humidity and air temperature at the standard reference height of 10 m, including their values for a neutral atmosphere approximation. The software tool can also be used if the meteorological data are collected over land, but the results will have some additional inaccuracies.

Specific application

Some example output for incoming and outgoing long-wave radiation and incoming and reflected short-wave radiation for 2004, calculated from data taken on a monitoring buoy at Esthwaite Water, UK, are shown in Figure 1. These data were calculated using LHFA, downloaded from the web and subsequently read into an excel file.

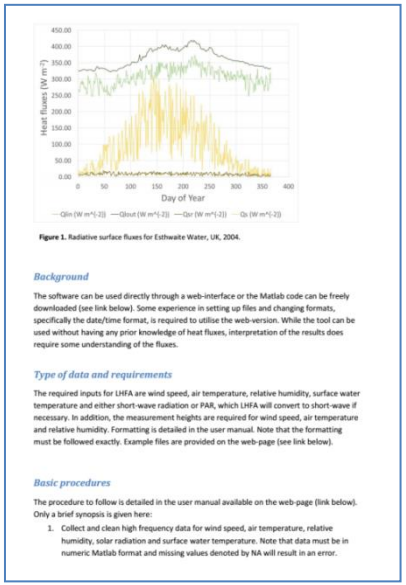


Figure 1. Radiative surface fluxes for Esthwaite Water, UK, 2004.

Background

The software can be used directly through a web-interface or the Matlab code can be freely downloaded (see link below). Some experience in setting up files and changing formats, specifically the date/time format, is required to utilise the web-version. While the tool can be used without having any prior knowledge of heat fluxes, interpretation of the results does require some understanding of the fluxes.

Type of data and requirements

The required inputs for LHFA are wind speed, air temperature, relative humidity, surface water temperature and either short-wave radiation or PAR, which LHFA will convert to short-wave if necessary. In addition, the measurement heights are required for wind speed, air temperature and relative humidity. Formatting is detailed in the user manual. Note that the formatting must be followed exactly. Example files are provided on the web-page (see link below).

Basic procedures

The procedure to follow is detailed in the user manual available on the web-page (link below). Only a brief synopsis is given here:

1. Collect and clean high frequency data for wind speed, air temperature, relative humidity, solar radiation and surface water temperature. Note that data must be in numeric Matlab format and missing values denoted by NA will result in an error.

QA/QC and data visualization

NETLAKE toolbox for the analysis of high-frequency data from lakes

NETLAKE

cost EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

Factsheet #1

Data handling: cleaning and quality control

Elvira de Eyto and Don Pierson

NETLAKE toolbox for the analysis of high-frequency data from lakes

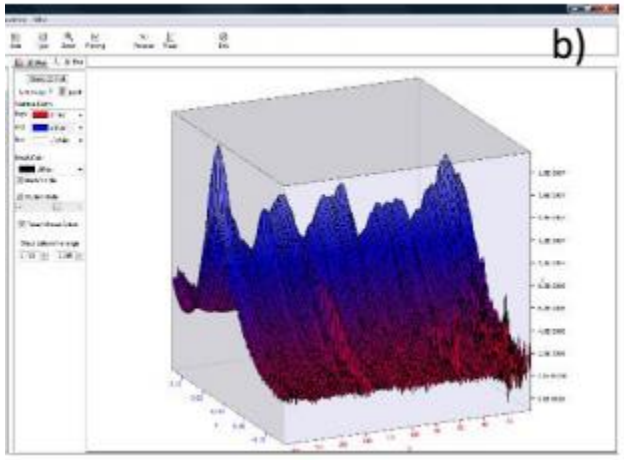
NETLAKE

cost EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

Factsheet #5

High frequency data treatment and visualization with ECDSOFT and OnLineMonitor

Dario Omanović and Ivanka Pižeta



NETLAKE toolbox for the analysis of high-frequency data from lakes

NETLAKE

cost EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

Factsheet #7

Knowledge Discovery in Databases - Data Mining

Ivanka Pižeta

Modelling tools

NETLAKE toolbox for the analysis of high-frequency data from lakes

NETLAKE

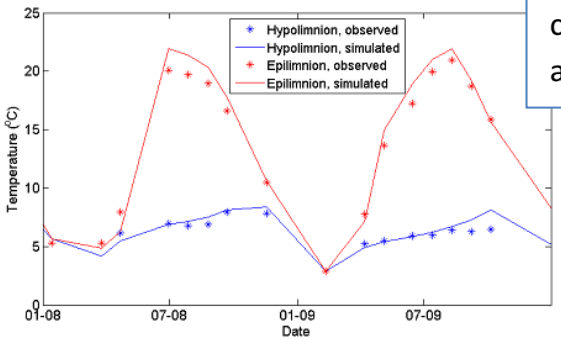
cost
EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

Factsheet #3

The General Lake Model (GLM)

Marieke Frassl, Michael Weber and Louise Bruce

The General Lake Model (GLM) is a one-dimensional hydrodynamics model. Hydrodynamic models describe the thermal properties and the mixing dynamics in water bodies. Based on inflow and outflow data, as well as meteorological data, GLM calculates a water and energy balance resulting in vertical profiles of temperature, salinity and density over time. As a one-dimensional model, GLM simulates the vertical profiles at one spatial point in the lake. Effects of ice cover on thermal properties and mixing of the lake can be included. GLM can also be coupled to biogeochemical models (e.g. AED, FABM), and therefore serves as the basis for models simulating the biological and chemical parameters in the water column. Data from monitoring stations are used as input data and to calibrate and validate the lake model. In combination with the observed data, GLM can be used to explore the role that stratification and vertical mixing play on the dynamics of lakes.



Modelling tools

NETLAKE toolbox for the analysis of high-frequency data from lakes

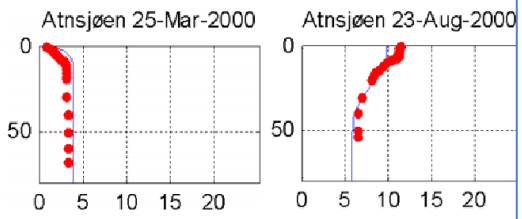
NETLAKE

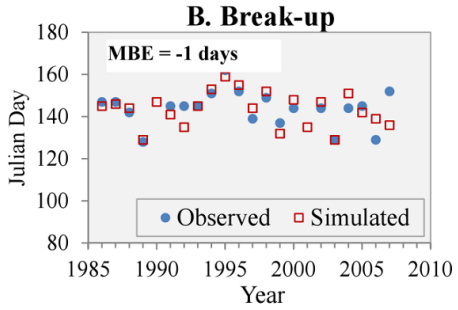
cost EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

Factsheet #6

Lake stratification and ice phenology: Modelling with MyLake

Raoul-Marie Couture and Koji Tominaga





Objective

Lake modelling is a thriving field of research, and many modelling tools are now available to the researchers (see Janssen et al. 2015). Prospective users of a model will make a choice based, amongst others, on the desired level of complexity and their preferred scientific programming environment. Here we describe the MyLake lake model, a simple one-dimensional (1D) daily time-step model that can be used to simulate seasonal changes in ice coverage in lakes. This model is aimed at researchers who prefer to use Matlab/Octave language for scientific computing applications. This factsheet describes briefly how to set-up the MyLake lake model in order to simulate thermal stratification and ice phenology in a lake.

Calculation of specific lake attributes

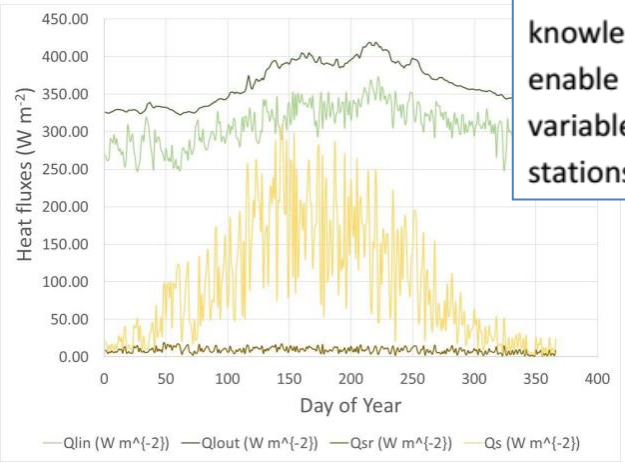
NETLAKE toolbox for the analysis of high-frequency data from lakes

Factsheet #2

Lake Heat Flux Analyzer (LHFA)

Ian Jones

Lake thermal and mixing properties are mainly driven by fluxes of heat and wind mixing at the surface of a lake. There are several different types of heat fluxes. With the right equipment these can be measured, but such equipment can be expensive and requires expertise to deploy. As an alternative to direct measurement most of these fluxes are often calculated with established methods using the meteorological variables commonly measured by automatic lake monitoring stations. These methods can be quite detailed and require some specialist knowledge to execute. The software tool, Lake Heat Flux Analyzer (LHFA), has been written to enable the calculation of these fluxes, and related terms, from standard meteorological variables. It has been specifically written for those using data from high resolution monitoring stations on lakes. The principal fluxes calculated are Q_{sr}, the reflected short-wave radiation;



Calculation of specific lake attributes

NETLAKE toolbox for the analysis of high-frequency data from lakes

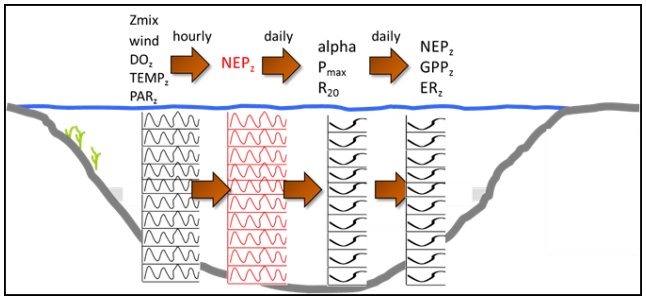
NETLAKE

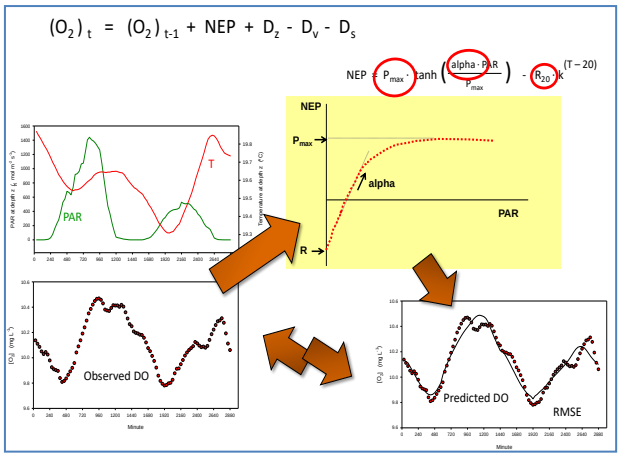
cost EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

Factsheet #9

Determination of whole-column metabolism from profiling data

Biel Obrador, Jesper Christensen and Peter A. Staehr





NETLAKE toolbox for the analysis of high-frequency data from lakes

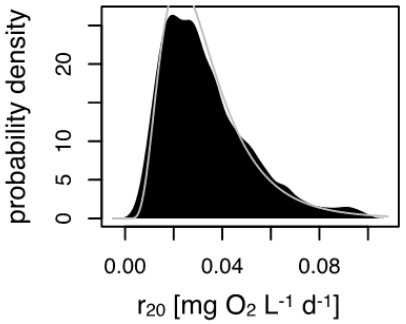
NETLAKE

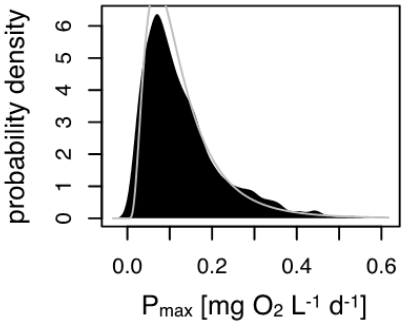
cost EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

Factsheet #8

Bayesian calibration of mechanistic models of lake metabolism

Mark Honti





Calculation of specific lake attributes

NETLAKE toolbox for the analysis of high-frequency data from lakes

NETLAKE

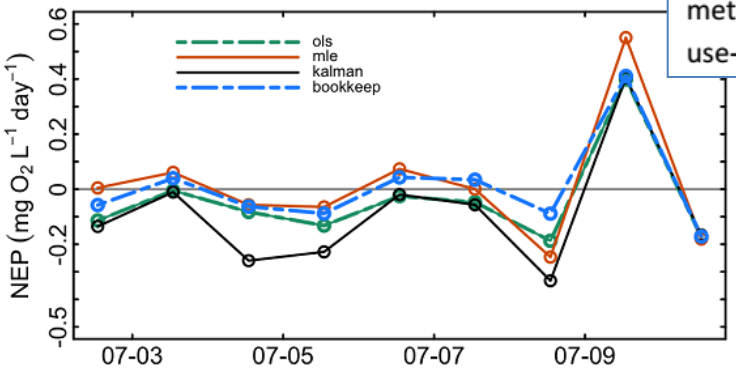
cost
EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

Factsheet #4

Lake Metabolizer

R. Iestyn Woolway

Lake Metabolizer is an Rpackage for estimating lake metabolism and related terms from data collected by high frequency, *in situ* lake monitoring stations with relative ease. The package can be used to calculate lake metabolism using five different methods: bookkeeping, ordinary least squares, maximum likelihood, Kalman filter, and Bayesian (Table 1). For further information of the differences between the metabolism models, see Winslow et al. (*in press*) and Honti (2016). In addition, each of these five methods can be combined with one of seven models for computing the gas transfer coefficient, which influences the rate of gas exchange at the air-water interface. Lake Metabolizer also includes a number of functions that compute conversions and calculations that are commonly applied to raw data prior to estimating lake metabolism (e.g. optical conversion models). This package contains example data, example use-cases, and function documentation.



Statistical techniques

NETLAKE toolbox for the analysis of high-frequency data from lakes

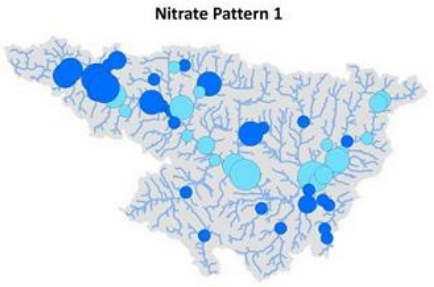
NETLAKE

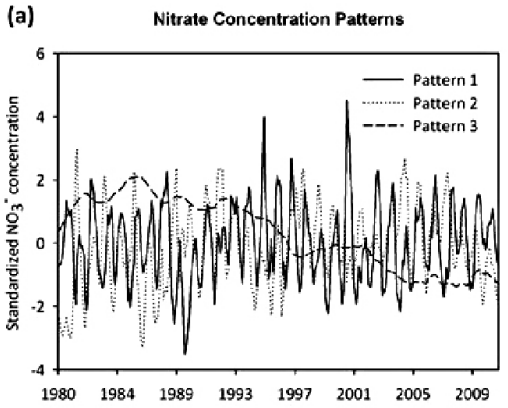
cost EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

Factsheet #10

Pattern detection using Dynamic Factor Analysis (DFA)

Rosana Aguilera and Rafael Marcé





Objective

One of the main applications of time series analysis is the identification of trends and cyclic patterns in the data. Many trend detection and frequency decomposition analyses already exist for those purposes, particularly to address single time series. However, classical methodologies are not particularly well suited to cope with multivariate problems. Dynamic Factor Analysis (DFA) decomposes a collection of time-series into common patterns and associated error terms (Zuur *et al.* 2003a). Broadly speaking, this method resembles performing Principal Component Analysis (PCA) but it is specifically designed for time-series. The end-product is a collection of patterns shared by all time series, the relative relevance of each pattern across time-series, and errors terms.

Statistical techniques

NETLAKE toolbox for the analysis of high-frequency data from lakes

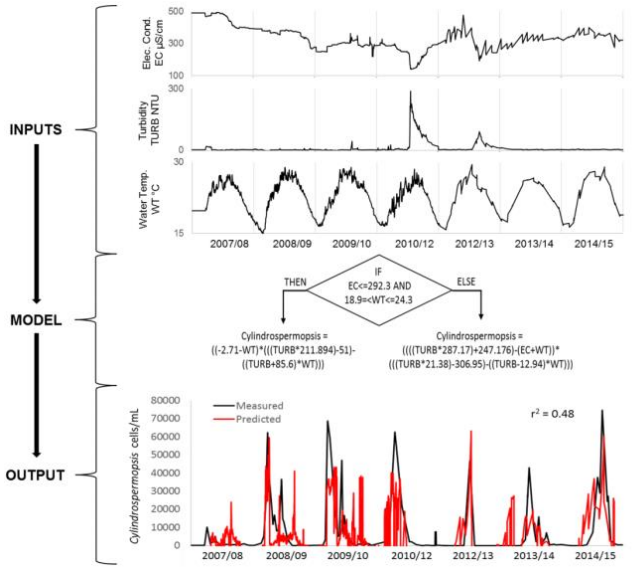
NETLAKE

cost
EUROPEAN COOPERATION
IN SCIENCE AND TECHNOLOGY

Factsheet #11

Inferential modelling of time series by evolutionary computation

Friedrich Recknagel and Ilia Ostrovsky

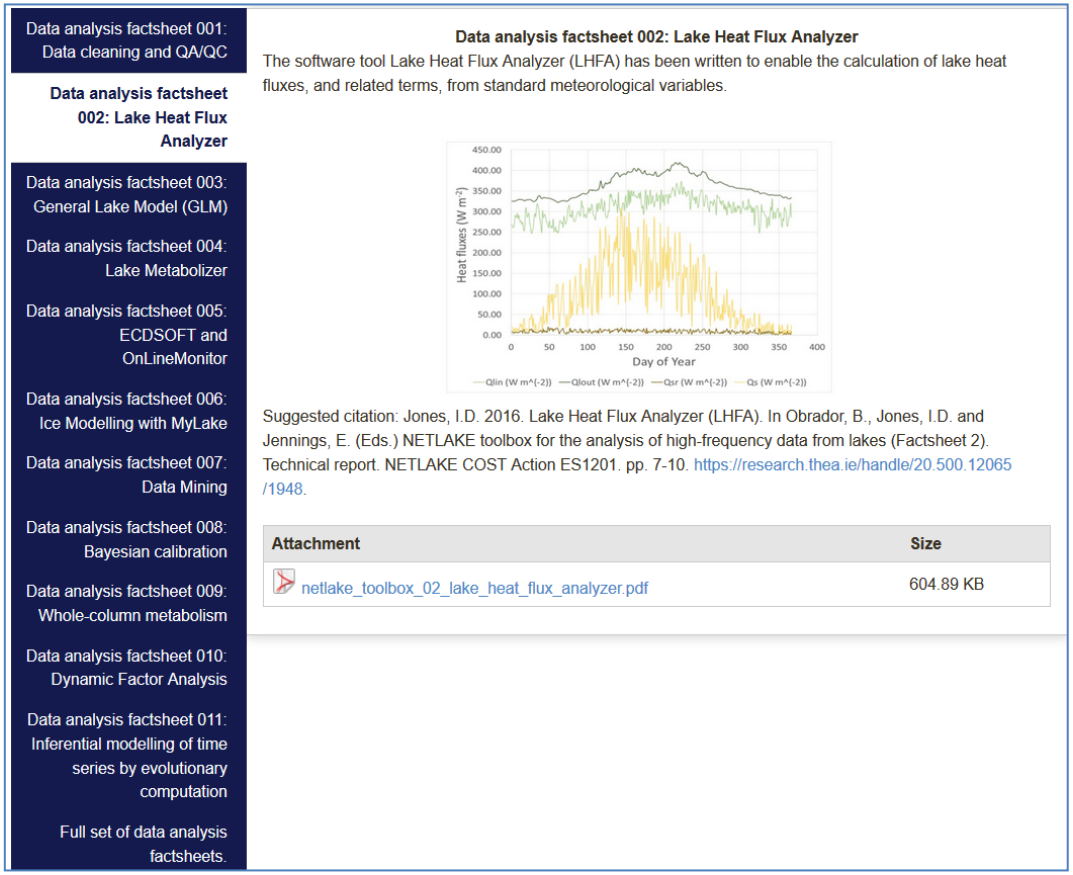


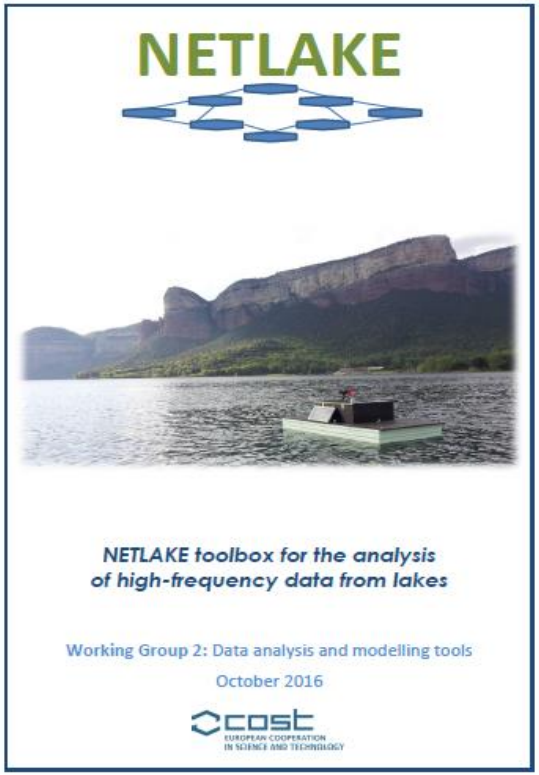
The hybrid evolutionary algorithm (HEA) has been designed: 1) to represent and forecast multivariate relationships between environmental conditions and population densities by inferential (IF-THEN-ELSE) models, and 2) to quantify 'tipping points' for population outbreaks by IF-conditions (Figure 1). During the course of hundreds of iterations, HEA discovers the 'best-fitting' model after optimising model structures by genetic programming and model parameters by differential evolution towards the lowest RMSE and highest R^2 (Cao et al. 2013).

NETLAKE toolbox

Individual factsheets and complete booklet are **available online** as downloadable pdfs at the NETLAKE webpage

www.dkit.ie/netlake





The NETLAKE toolbox is not intended to be an exhaustive list, but rather a **useful introduction** to some of the **many tools and techniques** currently used by specialists to process, analyse and model high-frequency data from lakes.

A mixing event in an elevator, after discussing the TOOLBOX

Riga, 2016