Globolakes
WP4: Data Integration and Uncertainty Budgets

January 2014

GloboLakes

Global Observatory of Lake Responses to Environmental Change

CEH Centre for Ecology & Hydrology
NATURAL ENVIRONMENT RESEARCH COUNCIL
UNIVERSITY OF DUNDEE
University of Reading
University of Glasgow
PML Plymouth Marine Laboratory
UNIVERSITY OF STIRLING

WP4- Integrate data and construct uncertainty budgets

Aim

Construct uncertainty budgets for each of the different data sources to incorporate in the EO calibration.

Objectives

4.1 Estimation of errors & uncertainties associated with in-situ data sets of lake state & condition for EO validation.

4.2 Full assessment of errors & uncertainties associated with the generated EO products as inputs into WPs 5 & 6.

WP4- Integrate data and construct uncertainty budgets

- Led by University of Stirling and contributions from all.
- The consortium met in January 2013 to discuss what was already known about the nature and type of errors which can arise for each data source within Globolakes.
- A working report from the meeting collates the discussion from the meeting, along with terminology to be used by the consortium partners for this work.

Globolakes
WP4: Statistical Modelling for Uncertainty Assessment

Ruth Haggarty, Claire Miller, Marian Scott

January 2014

GloboLakes

Global Observatory of Lake Responses to Environmental Change

Centre for Ecology & Hydrology
NATURAL ENVIRONMENT RESEARCH COUNCIL

University of Reading

University of Glasgow

PML | Plymouth Marine Laboratory

UNIVERSITY OF STIRLING

Statistical modelling for uncertainty assessment

For a small subset of lakes:

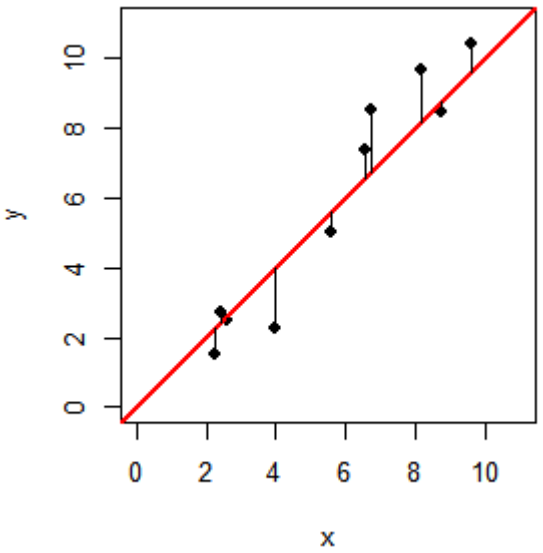
1. **What effect will errors in the In-situ data have on the validation of the EO products? What is the scale of the errors?**

2. **How do we match In-situ data and EO data at different spatiotemporal scales?**

   Combining information from observed point-referenced monitoring data and EO products (hierarchical models, downscaling co-kriging).

Errors in Variables Approaches

In standard regression the line of best fit is estimated by minimising the sum of the squared vertical distances between the observed responses and the responses predicted by the linear approximation.



Here, the assumption is made that there is no measurement error associated with x, and only y has been measured with some degree of imprecision.

If it is thought that there is measurement error associated with both variables and this is not taken into account in some way, the parameters estimated to describe the relationship will potentially be biased.

The results of any hypothesis tests on the regression parameters estimated may therefore be invalid.

Errors in variables models take into account variability in both variables.

Regression Models

6 Different regression approaches have been considered:

1. **Ordinary least squares** – standard regression approach, assumes error is only present in the remote sensing data.

2. **Weighted least squares** – as with OLS, however weights are used in an attempt to account for non-constant variability in the errors.

3. **Deming regression** – accounts for error in both variables and can be applied where there is a higher level of error in one measurement than the other.

4. **Weighted Deming regression** – as Deming regression however weights are used in an attempt to account for non-constant variability in the errors.

5. **Geometric Mean Regression** – accounts for error in both variables, assumes same error variance for both measurements.

6. **Passing Bablok Regression** – non-parametric approach which is robust to outliers and makes no distributional assumptions on errors. All other approaches assume normally distributed errors.

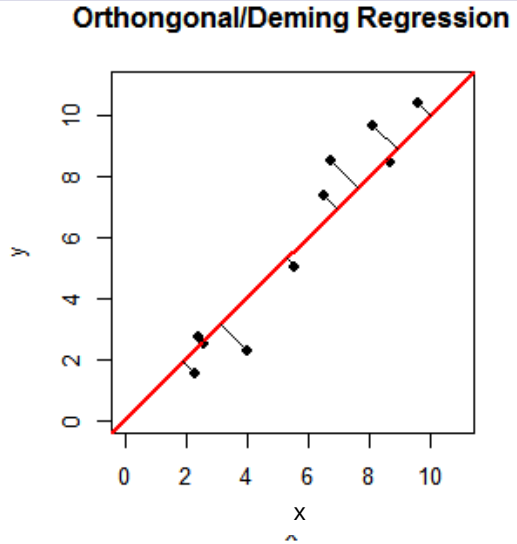
Orthogonal/Deming Regression

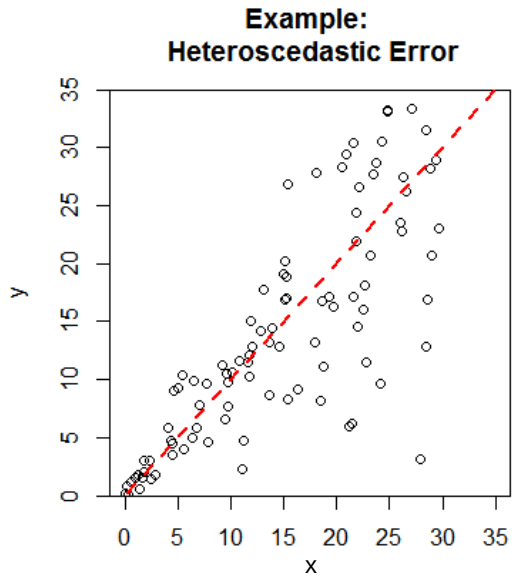
- **Orthogonal Regression** allows for imprecision in both x and y by choosing the line that minimises the sum of squared differences from the observations to the line that are in a direction at right angles to the line.

- **Deming regression** extends orthogonal regression for minimisation of distances at angles other than 90 degrees.

- Deming regression incorporates information from the ratio of error variability in the in-situ and the remote sensing data

- **Weighted Deming regression** can be applied to additionally account for heteroscedasticity in the error (variability changes as mean changes)





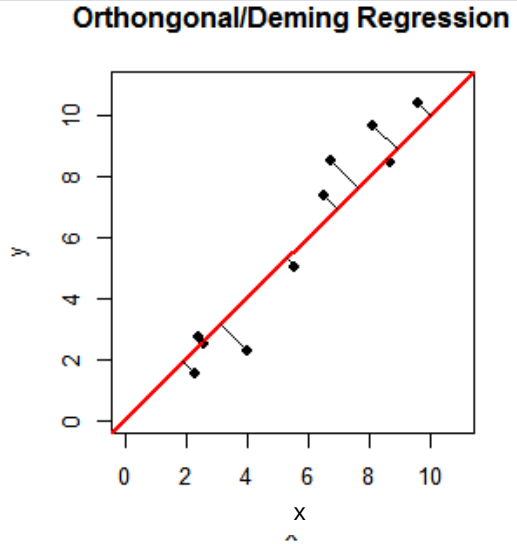
Orthogonal/Deming Regression

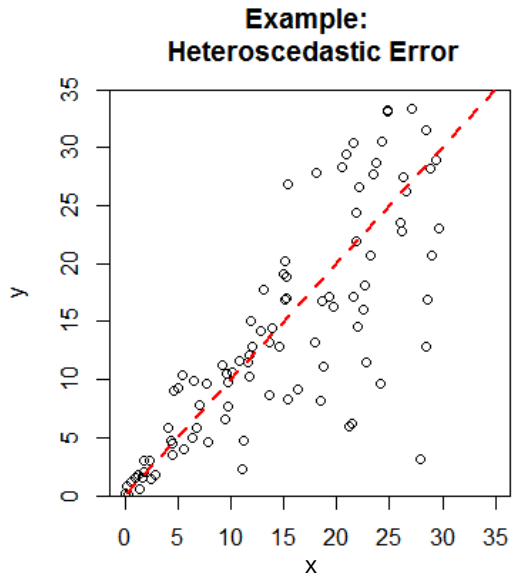
- **Orthogonal Regression** allows for imprecision in both x and y by choosing the line that minimises the sum of squared differences from the observations to the line that are in a direction at right angles to the line.

- **Deming regression** extends orthogonal regression for minimisation of distances at angles other than 90 degrees.

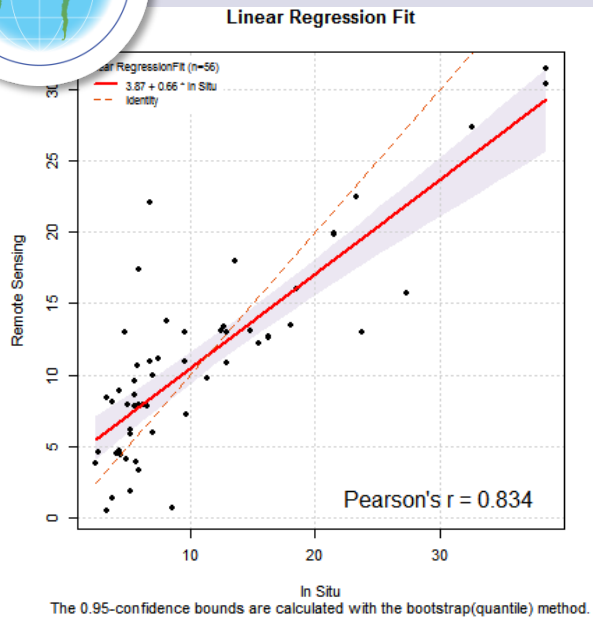
- Deming regression incorporates information from the ratio of error variability in the in-situ and the remote sensing data

- **Weighted Deming regression** can be applied to additionally account for heteroscedasticity in the error (variability changes as mean changes)

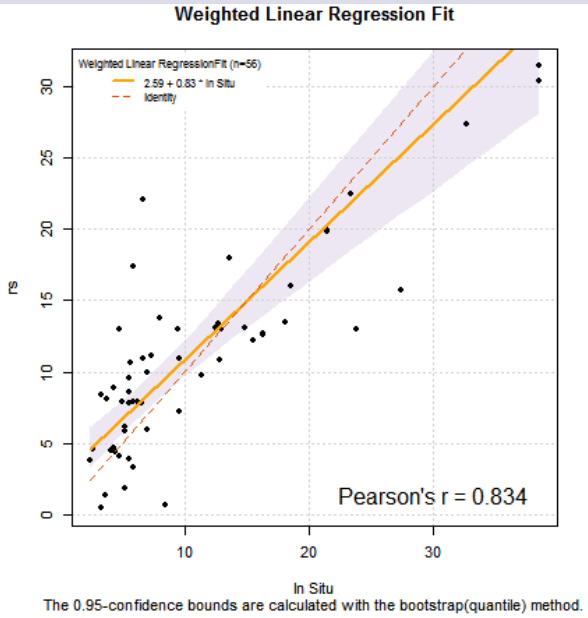




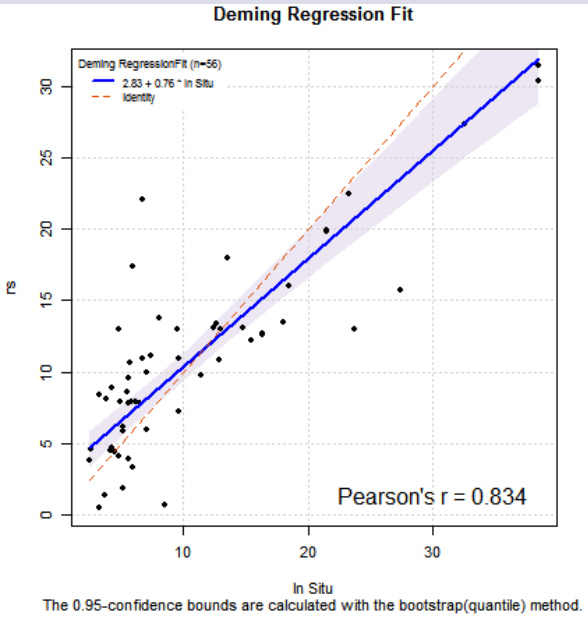
MCI Chl Fitted Regression Lines



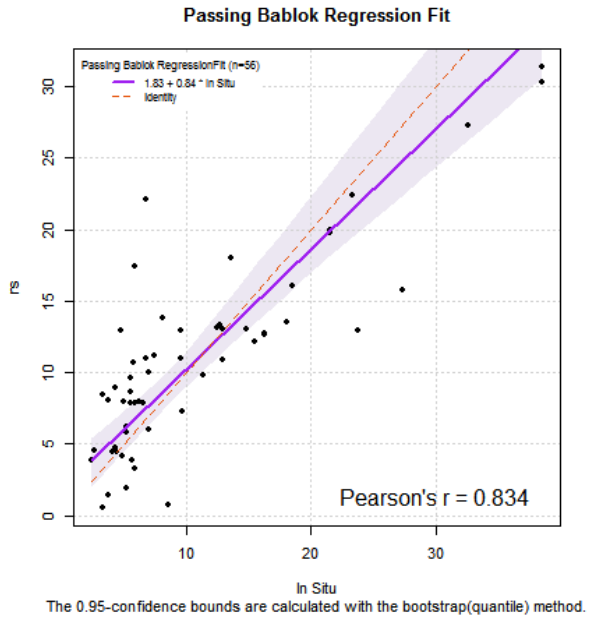
The 0.95-confidence bounds are calculated with the bootstrap(quantile) method.



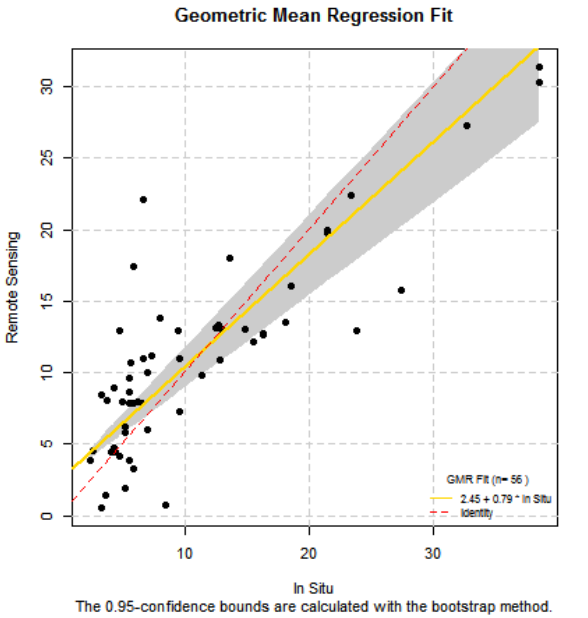
The 0.95-confidence bounds are calculated with the bootstrap(quantile) method.



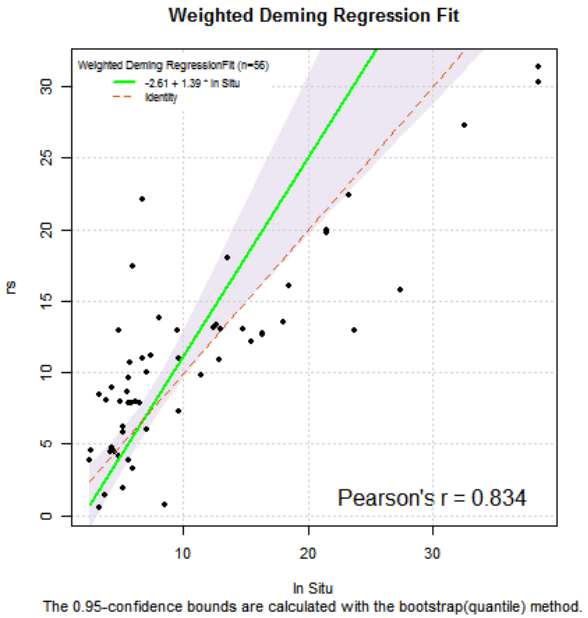
The 0.95-confidence bounds are calculated with the bootstrap(quantile) method.



The 0.95-confidence bounds are calculated with the bootstrap(quantile) method.

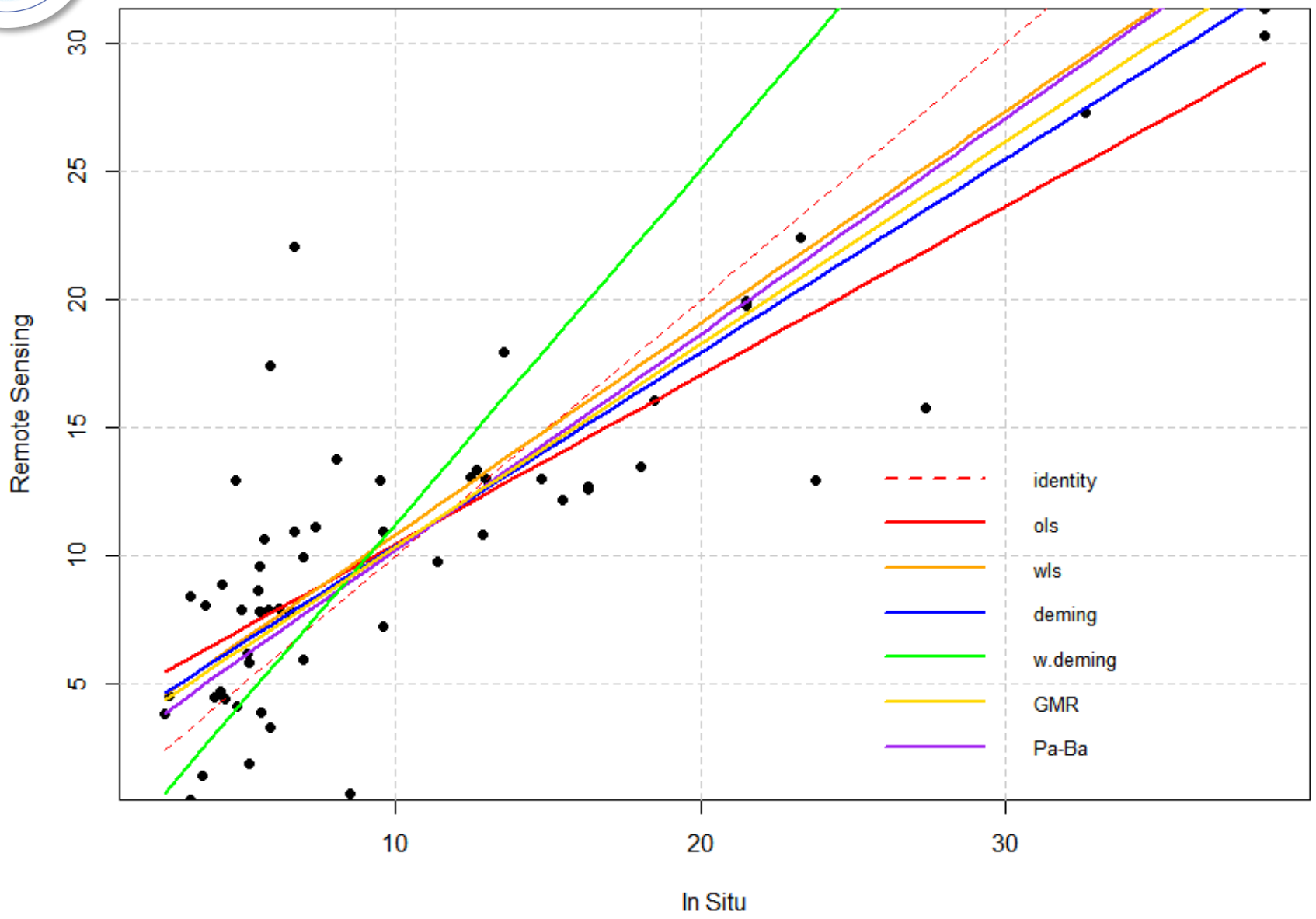


The 0.95-confidence bounds are calculated with the bootstrap method.

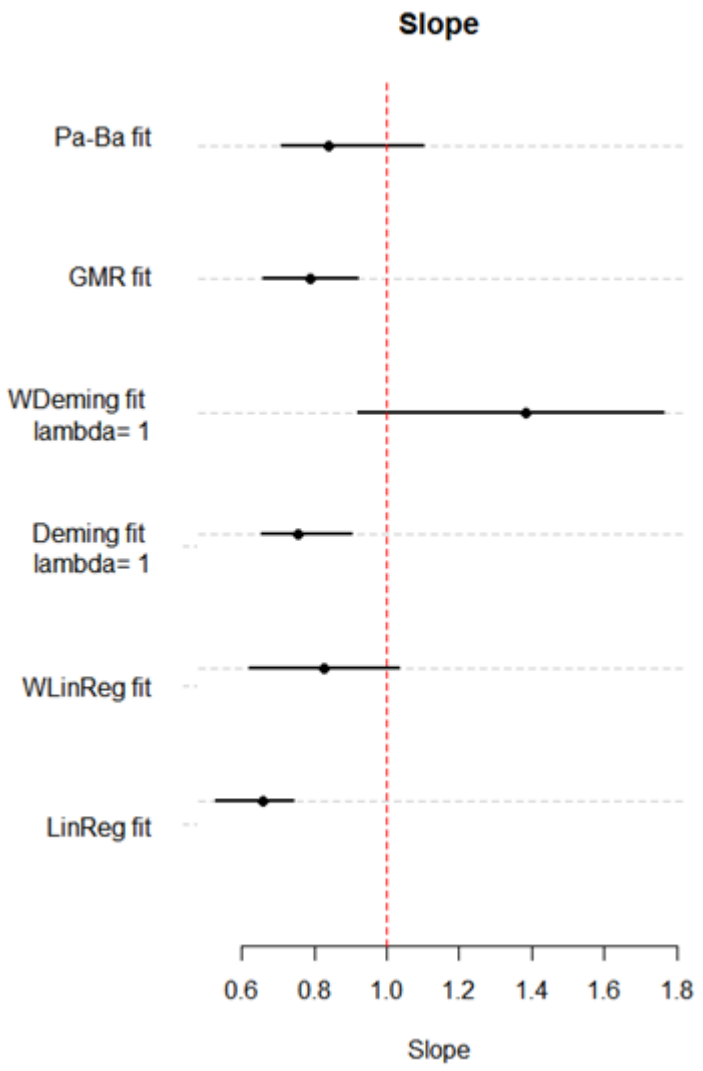


The 0.95-confidence bounds are calculated with the bootstrap(quantile) method.

MCI Chl Fitted Regression Lines



- Plot shows confidence interval for the slope under each regression approach.
- In this case P-B, Weighted Deming regression and Weighted Linear regression all identify that the identity relationship is present - these are the approaches which take into account non constant variance.
- The widths of confidence intervals are quite different.



Simulation Study

- A simulation study was designed to investigate the ability of each of the approaches to detect a y=x relationship, and a linear relationship, in the presence of a range of different error structures.

The study considered the effects of :

- Varying numbers of samples.
- Different quantities of error variability.
- Different ratios of error variability (λ denotes the error variance ratio).
- Different degrees of heteroscedasticity in the error terms.

Simulation Study

- For each simulation scenario the statistical power and size was computed:
  - **Power** is the probability of detecting the y=x relationship (or any statistically significant linear relationship) when it is present.
  - **Size** is the probability of detecting the y=x relationship (or any statistically significant linear relationship) when there is no underlying relationship present. Results for this are not presented here.
- 500 datasets, each with 100 samples were simulated for each scenario and the 6 regression models discussed were fitted to each dataset.

Simulation Scenarios

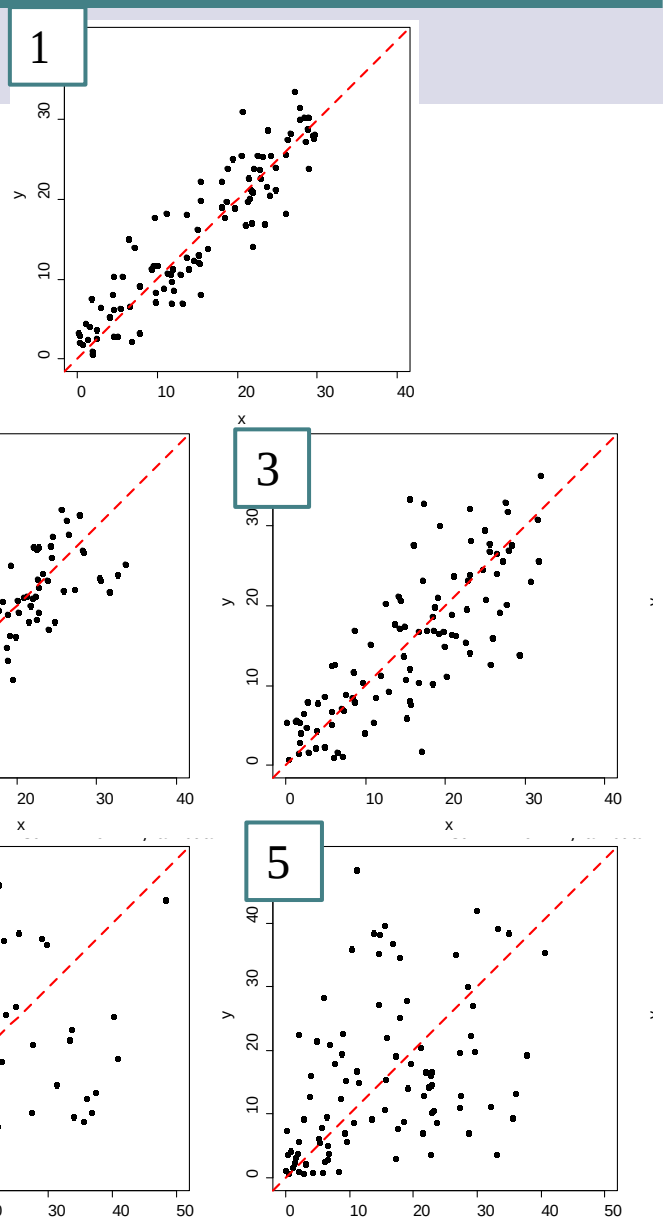
1. x with no error, y with $N(0,\sigma^2)$ error.

2. x and y with constant error, both $\sim N(0,\sigma^2)$.

3. x and y with constant error,
   x error $\sim N(0,\sigma^2)$
   y error $\sim N(0,2\sigma^2)$

4. x and y with heteroscedastic error, both equal.

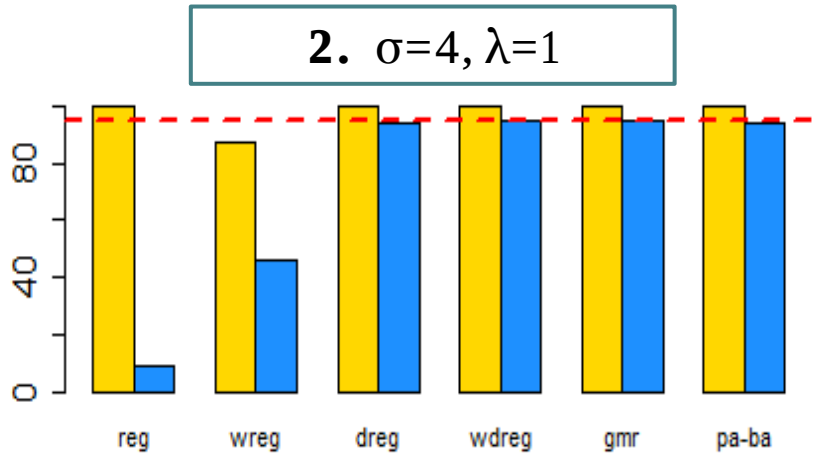
5. x and y with heteroscedastic error,
   y error = 2 * x error



Simulation Results: Example

Power Results for scenarios 2 and 3

- Red line corresponds to 95% power.
- Yellow bar is power to detect any linear relationship.
- Blue bar is power to detect true x=y relationship.
- Results for non-constant variance are not shown but for λ=1 weighted deming regression performs well in terms of the level of power achieved



Simulation Results: Summary

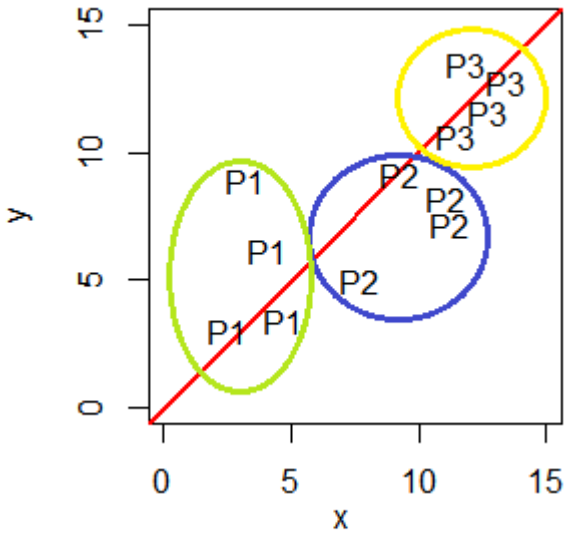
- When the underlying data have error in both variables the errors in variables approaches outperform the standard approaches in terms of detecting the true underlying relationship.
- Deming regression and Weighted Deming regression perform well when there are different error variance values for each variable ($\lambda \neq 1$).
- Weighted Deming regression performs well when there is non constant variance and the error variance ratio is equal to 1 ($\lambda = 1$).
- When there is both non constant variance and a error variance ratio (λ) which is not 1 the methods struggle to detect the underlying pattern.

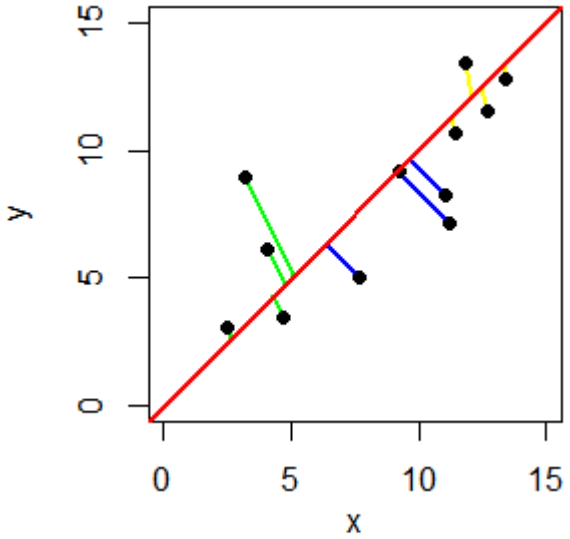
Extensions

It is unlikely matchups are considered at single lake.

We are developing *Modified Deming Regression*.

- Multiple error variance ratio values can be used.
- One error variance ratio for each lake.
- In practice this means regression line is obtained by minimising distance lines at different angles simultaneously.
- This approach assumes that the variability across lake is variable, but within lake is constant.
  - Further modifications to the weights can be used to account for changes in variability within lake, if necessary.





Summary

- A range of different errors in variables approaches have been considered (others are available).
- While the point estimates of the slope are often similar, there can be large discrepancies in the standard errors associated with these estimates.
- Weighted approaches enable us to account for heteroscedasticity and to combine information from multiple lakes.
- Further investigation is ongoing to estimate the size of the error ratios in real data. Additional data on both remote sensing values and in-situ values will be used in order to quantify and incorporate this information.
- Deming regression and geometric mean regression can both be extended beyond method comparison and can be used when there are multiple explanatory variables which are all subject to uncertainty.

Related Work

- MSc project exploring spatial variability at Lake Balaton.
  - Exploring in-situ data at 5 different sites.
  - Exploring ArcLake data, pixel variability.

- PhD project comparing spatiotemporal variability for large lakes.
  - Focussed work comparing lake surfaces.

...both in early stages!

References

- Deming, W. E. *Statistical adjustment of data*. New York: Wiley (1943)
- Draper, N.R., Smith, H. *Regression Analysis*. New York: Wiley (1998)
- Gillard, J., Iles, T. Methods of Fitting Straight Lines where Both Variables are Subject to Measurement Error. *Current Clinical Pharmacology* 4: 164-171. (2009)
- Linnet, K. Estimation of the linear relationship between the measurements of two methods with proportional errors. *Statistics in Medicine* 9;12: 1463- 1473 (1990)
- Linnet, K. Evaluation of regression procedures for method comparison studies. *Clinical Chemistry* 39; 3 : 424–432 (1993)
- Ricker, W. E. Linear regression in fisheries research. *Journal of the Fisheries Research Board of Canada* 30: 409-434 (1973)