

# **The Molten Diatom Database and Transfer Functions**

User Guide

Version 1.0  
January 2004

Steve Juggins  
School of Geography, Politics and Sociology  
University of Newcastle  
Newcastle upon Tyne  
NE1 7RU, UK  
[Stephen.Juggins@ncl.ac.uk](mailto:Stephen.Juggins@ncl.ac.uk)



## Contents

Contents .....	2
1 Introduction .....	3
2 MOLTEN datasets .....	3
3 Diatom taxonomy.....	5
4 Environmental reconstructions .....	6
4.1 Which training set should I use? .....	6
4.2 Evaluating the reconstruction.....	6
4.3 Data preparation.....	7
4.4 Transfer function description.....	7
4.5 Reconstruction software (ERNIE).....	9
4.6 Performing a TN reconstruction.....	9
5 References.....	13
Appendix 1 List of MOLTEN dataset / transfer function contributors .....	14



## 1. Introduction

The MOLTEN diatom database is an information system designed to enhance the application of diatom analysis to problems of coastal eutrophication. It has been produced by combining and harmonising data from a series of smaller datasets from Denmark, Sweden, Finland and the Netherlands and includes electronic images of diatoms, new training sets and transfer functions for inferring total nitrogen from fossil diatom assemblages, and software for applying the transfer functions. It is the result of a three-year collaboration between diatom taxonomists, palaeolimnologists, statisticians and database experts. MOLTEN was funded by the European Commission under the V framework Environment and Climate Programme (Grant No. EVK3-CT-2000-00031).

The MOLTEN system comprises both an interactive web site for browsing site and diatom taxonomic information (<http://craticula.ncl.ac.uk:8000/Molten/jsp>) and stand-alone software for inferring or hindcasting total nitrogen concentrations from fossil diatom assemblages. This guide describes the web-based information system and stand-alone software. A review of the transfer function approach to environmental reconstruction can be found in Birks (1995), and ter Braak & Juggins (1993).

## 2. MOLTEN datasets

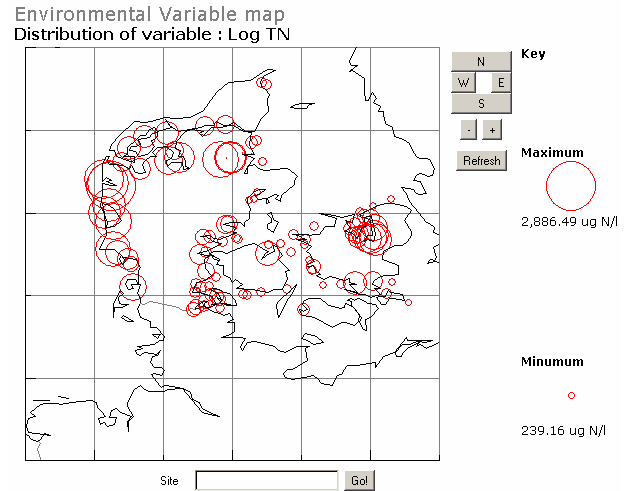
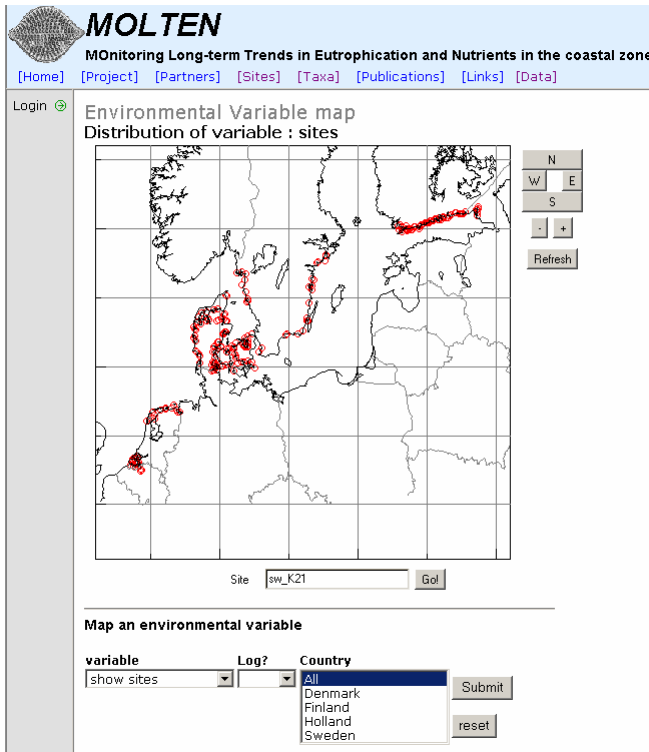
The main resource of the MOLTEN system are four regional diatom training sets, consisting of modern surface sediment diatom samples and associated water chemistry data. The datasets are from Denmark (N=91), Sweden (N=35), Finland (N=55) and the Netherlands (N=25). Table 1 lists summary chemistry for the four datasets.

**Table 1. Summary chemistry for the four MOLTEN datasets**

Variable	Dataset	N	Minimum	Maximum	Mean	Median	SD
TN ( $\mu\text{g l}^{-1}$ )	Denmark	91	239.2	3402.0	779.6	549.5	584.5
	Holland	25	596.9	3890.3	1409.2	1178.3	929.2
	Finland	39	381.3	1245.0	701.8	712.5	169.8
	Sweden	35	260.2	1335.0	474.8	395.6	228.8
Depth (m)	Denmark	91	1.1	40.0	13.2	10.1	9.1
	Holland	25	3.0	50.5	17.3	15.1	12.7
	Finland	55	0.7	29.8	4.9	3.2	5.0
	Sweden	35	0.5	74.0	22.3	17.0	19.1
Salinity ( $\text{g l}^{-1}$ )	Denmark	91	2.5	31.1	18.8	18.0	5.7
	Holland	25	14.7	31.6	26.6	29.0	5.0
	Finland	55	0.6	6.3	4.7	4.9	1.2
	Sweden	35	0.9	26.0	9.8	6.6	7.4
TP ( $\mu\text{g l}^{-1}$ )	Denmark	91	15.5	442.6	67.4	47.5	78.2
	Holland	25	55.9	437.0	135.5	86.5	97.9
	Finland	55	14.3	82.1	34.7	33.4	12.8
	Sweden	35	17.0	114.7	32.3	26.8	19.3
NO <sub>x</sub> _N ( $\mu\text{g l}^{-1}$ )	Denmark	91	23.5	2761.0	306.8	139.6	426.3
	Holland	25	183.8	3208.9	781.6	402.0	778.7
	Finland	55	5.0	344.1	34.2	12.3	59.3
	Sweden	35	27.4	428.4	115.2	76.3	96.2
NH <sub>4</sub> _N ( $\mu\text{g l}^{-1}$ )	Denmark	91	10.3	189.1	49.3	46.1	33.0
	Holland	25	54.3	223.3	99.2	91.9	36.4
	Finland	55	5.8	1566.7	62.2	19.7	208.4
	Sweden	35	2.0	155.2	24.7	15.6	27.7



The MOLTEN website has a java applet for plotting these datasets and environmental characteristics.



The figures above show the location of the four regional datasets and the distribution of total nitrogen in the Danish training set.

Clicking on an individual sample on the map lists detailed site, chemical and diatom information for that sample:

Sample information

Site Id: NJYM4

Original Site Id	NJYM4
Site Name	Manager Fjord
Country	Dk
Latitude	56° 38' 0"
Longitude	9° 50' 0"
Depth	6.5

Summary of environmental data:

Variable	Units	Value
temperature		9.2
salinity		15.4
PO4_P		67.7
TP		108.6
NO3_N		1023.0
NH4_N		189.1
TN		1840.1
SiO3_Si		3257.6
chlorophyll_a		9.1
secchi		4.2

Top 10 taxa:

Taxa
<a href="#">Thalassiosira sp. 1A</a>
<a href="#">Cyclotella choctawhatcheeana</a>
<a href="#">Opephora mutabilis</a>
<a href="#">Fragilaria elliptica agg.</a>
<a href="#">Staurosira construens var. venter</a>
<a href="#">Fallacia sp. 4A</a>
<a href="#">Cyclotella meneghiniana</a>
<a href="#">Staurosirella leptostauron</a>
<a href="#">Pseudostaurosira brevistriata</a>
<a href="#">Bacillaria paxillifer</a>



### 3. Diatom taxonomy

Within MOLTEN the taxonomy of the four regional datasets has been harmonised to ensure a consistent approach to identification and nomenclature. To use the MOLTEN transfer functions any diatom counts from fossil sediments must obviously use the same nomenclature and identification criteria. The MOLTEN website containing a list of all diatom data recorded in the MOLTEN datasets, along with its a unique 6-character code used to identify that taxon in the reconstruction software. A full list of MOLTEN taxa and their codes can be found in Appendix 1 of this document. The website also contains images of key taxa to aid in the harmonisation of sediment core diatoms counts with the MOLTEN taxonomy.

The first taxon tool on the MOLTEN website is the Genera browse tool.

#### Genera browse

This page lists all genera which have species with a maximum abundance >1% and found in more than one site. Log in for full access to the database. After publication this restriction will be removed. Click on a name to view a list of species.

Genus	N	Genus	N	Genus	N
Achnanthes	11	Diploneis	9	Pinnularia	3
Achnanthisdium	1	Encyonema	1	Plagiogramma	1
Actinocyclus	2	Encyonopsis	1	Plagiogrammopsis	3
Actinoptychus	1	Entomoneis	2	Plagiotropis	1
Amphora	14	Epithemia	3	Planothidium	2
Anaulus	1	Eunotia	1	Pleurosigma	1
Astartiella	1	Fallacia	11	Porosira	1
Asterionella	1	Fragilaria	12	Psammothidium	1
Aulacoseira	6	Fragilariforma	1	Pseudo-Nitzschia	1
Bacillaria	1	Fragilariopsis	1	Pseudostaurosira	1
Berkeleya	1	Frustulia	1	Reimeria	1
Biremis	2	Glyphodesmis	2	Rhabdonema	1
Brachysira	2	Gomphonemopsis	3	Rhaphoneis	1
Brockmanniella	1	Grammatophora	1	Rhoicosphenia	1
Campylodiscus	1	Gyrosigma	8	Rhopalodia	2
Catenula	1	Haslea	1	Stauroneis	1
Cavinula	1	Hyalodiscus	1	Staurosira	5
Cerataulina	1	Karayevia	2	Staurosirella	3
Cocconeis	14	Martyana	3	Stephanodiscus	5
Coccinodiscus	2	Mastogloia	3	Surirella	1
Ctenophora	1	Melosira	5	Tabularia	1
Cyclostephanos	1	Minidiscus	1	Tabularia	1
Cyclotella	10	Navicula	55	Thalassionema	1
Cymatosira	1	Nitzschia	23	Thalassiosira	22
Cymbella	2	Odontella	2	Trachyneis	1
Delphineis	2	Opephora	7	Tryblionella	6
Diatoma	2	Paralia	1	Unknown	12
Dickieia	1	Pauliella	1		
Dimeregramma	1	Petroneis	1		

Clicking on a name lists members of that genus found in MOLTEN samples:

List of taxa in *Cyclotella*

View another genus

Taxon code	Taxon name	N	Max %
CycAtto	<i>Cyclotella atomus</i>	55	35.6
CycCho	<i>Cyclotella choctawhatcheeana</i>	165	76.0
CycCom	<i>Cyclotella comensis</i>	14	6.0
CycMen	<i>Cyclotella meneghiniana</i>	69	10.0
CycPse	<i>Cyclotella pseudostelligera</i>	16	1.9
CycRad	<i>Cyclotella radiosa</i>	25	2.1
CycRos	<i>Cyclotella rossii</i>	10	1.3
CycSte	<i>Cyclotella stelligera</i>	17	1.5
CycStr	<i>Cyclotella striata</i>	55	2.7
CycStrAG	<i>Cyclotella striata agg.</i>	46	10.6

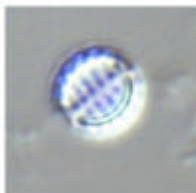
Total of 10 taxa matching Cyclotella in database.

#### Taxon Information

##### *Fragilaria elliptica* agg.

Informal authority: Authority not known

Molten taxon code: FrElIAG



name: cf. *Staurosira sopotensis*1  
Site: K15  
By: E

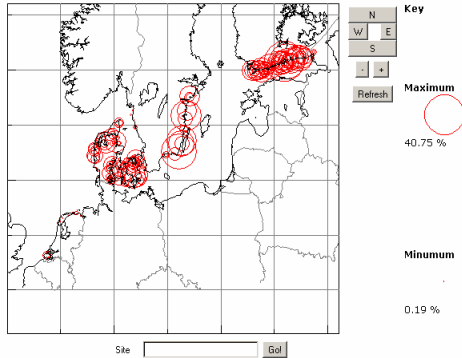
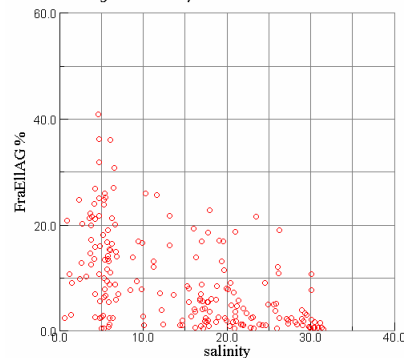
#### Distribution in Molten sites: top 10 sites

Dataset	Site
fr	fr_12
fr	fr_11
sw	sw_K18
sw	sw_S10a
sw	sw_K3v
sw	sw_K15
fr	fr_4
fr	fr_25
Dk	STR82008
Dk	STR13005

Draw a map to show the distribution of this taxon.

Selecting a taxon name in the species list (above right) lists taxonomic information for that taxon, including authority, and for key taxa, an image.

This page also lists the 10 MOLTEN samples which contain this taxon in the greatest relative abundance.

Taxon Distribution Map  
Distribution : FraEIIAGGraph of  
FraEIIAG % against salinity

Additional taxon tools allow the user to plot a distribution map of any MOLTEN taxon (above left), or to plot a taxon distribution along environmental or chemical gradients (above right).

## 4. Environmental reconstructions

### 4.1 Which training set should I use?

There is no single training set in MOLTEN that can be considered appropriate for application to all sediment core material. Statistical analysis of the separate Danish, Swedish and Finnish regional datasets and the combined training set demonstrates that total nitrogen explains a statistically significant part of the variance in the diatom distributional data. However, depth and salinity are also important in explaining variation in the diatom data, and examination of species responses shows that some taxa have different TN preferences in different datasets. It is unclear at present whether this is because of real ecological interactions with depth and salinity or due to numerical artefacts resulting from the rather disjunct environmental ranges of the different training sets.

Statistical analysis of the diatom and combined diatom / environmental data indicate that there are essentially two separate groups of diatom samples: a Western Baltic group comprising samples from Denmark and western Sweden and an Eastern Baltic group comprising samples from Finland and eastern Sweden. To this end we have merged the original MOLTEN regional datasets into the new Western and Eastern Baltic training sets and derived transfer functions for these training sets based on weighted-averaging partial least squares (WAPLS). We recommend that the Western Baltic training set be applied to cores collected from the Western Baltic and visa versa.

### 4.2 Evaluating the reconstruction

All MOLTEN transfer functions will provide a quantitative TN reconstruction. An important part of the transfer function approach to hindcasting TN concentrations is a thorough evaluation of the reliability of the reconstructed values. This can be done in three ways:

#### *1. Compare reconstructions using different training sets / transfer functions*

In addition to the Western and Eastern Baltic WAPLS transfer functions we also provide WAPLS and LWWA transfer functions developed using the combined dataset. We recommend that reconstructions also be performed using these transfer functions and the results compared. Given that some taxa have apparently different TN preferences in the different datasets we would expect the different transfer functions to produce different reconstructions. However, if downcore changes in diatom taxa are primarily driven by changes in TN concentrations then we would expect the three reconstructions to follow



similar trajectories, even if they differed in the absolute values of the hindcast TN concentrations. If the three methods produce widely divergent reconstructions this would suggest that the core contains key taxa whose distribution is not primarily related to TN, and as such, the reconstructions should be treated with caution.

## **2. Closest analogue analysis**

In addition to the WAPLS and LWVA reconstruction methods the MOLTEN software can also be used to find the closest modern analogue for each fossil sample using the modern analogue technique (MAT). MAT quantifies the dissimilarity between modern and fossil samples using squared chi-square distance. Precise interpretation of this analog measure is difficult but a rule of thumb suggests that any fossil sample that lies beyond a value of 70 has no close analogues in the training set (this estimate is based on the distribution of dissimilarities within the training set – see Juggins and Jones 1995 for a discussion). Thus reconstructions with a minDC value (see Section 4.6 below) greater than 70 should be treated with caution.

Output from the verification analysis can also help to identify no-analogue samples – that is, samples contain significant numbers of taxa absent from the training set.

## **3. Validation against historical TN concentration measurements**

The most robust evaluation of hindcast TN concentrations is to compare reconstructed values with historical time-series where these exist. Close agreement between hindcast and measured TN increases confidence in the reconstruction whereas divergence between the two sets of data warrants caution. However, close agreement during the monitoring period does not necessarily imply accurate reconstructions for earlier periods, especially if assemblages are very different. Conversely, divergence for the monitoring period does not necessarily imply reconstructions for earlier periods are in error.

## **4.3 Data preparation**

When using a MOLTEN transfer function it is essential that the taxonomy of the core matches that of the training set. In MOLTEN this means that the taxon codes used in the core and training set much match in terms of the codes themselves, and in the taxonomic concepts applied to the codes. In MOLTEN there are two ways to ensure agreement between training set and core taxonomy. The first and most straightforward is for the diatomist to use exactly the same codes in their core as are used in the training set. The MOLTEN website contains pages devoted to describing the taxonomy of the training sets to aid in this process. The second way to ensure compatibility is to use a conversion dictionary to convert between local codes and MOLTEN codes. This requires more work initially to setup the dictionary but is the most convenient if many cores are to be reconstructed as the same dictionary can be applied to each core.

MOLTEN software accepts core data in Cornell FULL or Condensed format. This format is used by other transfer function and ecological-statistical software. WinTran, a free program to convert from Excel, Tilia and other common spreadsheet and database formats into Cornell format is available for download from:

<http://www.campus.ncl.ac.uk/staff/Stephen.Juggins/>

## **4.4 Transfer function description**

MOLTEN contains three screened training sets / transfer functions for reconstructing total

nitrogen (Table 2).

**Table 2: MOLTEN training sets**

Training set	Number of samples	Total nitrogen range ( $\mu\text{g l}^{-1}$ )
Combined	189	239 – 3890
Western Baltic	101	239 – 2890
Eastern Baltic	63	263 – 1330

The combined datasets includes surface samples from all four regional training sets. The Western Baltic dataset includes samples from Denmark and the Swedish west coast, and the Eastern Baltic dataset contains samples from Finland the Swedish east coast.

The three training sets are used to develop 4 transfer functions. Three of these are developed using weighted averaging partial least squares (WAPLS). The fourth is developed using locally-weighted weighted-averaging (LWWA), a dynamic method that defines a local training set of 10 samples for each fossil sample, and uses this sub-set to generate a transfer function that is applied to each fossil sample in turn. The performance of the different transfer functions is listed in Table 3.

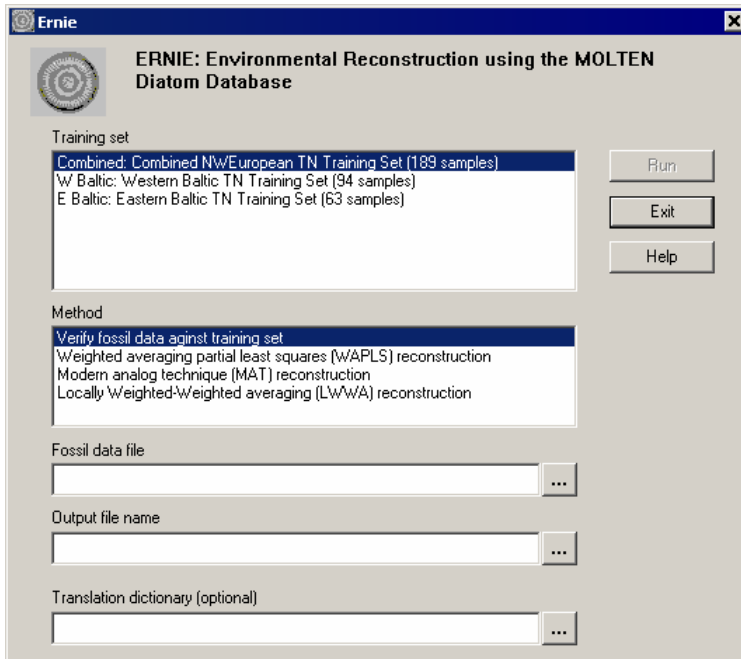
**Table 3: MOLTEN transfer functions showing transfer function performance measured by the root mean squared error of prediction under leave-one-out cross validation. Units are  $\log_{10}$  TN ( $\mu\text{g l}^{-1}$ ).**

Training set	Weighted averaging partial least squares (WAPLS)	Locally weighted weighted averaging (LWWA)
Combined	0.12	0.12
Western Baltic	0.14	NA
Eastern Baltic	0.10	NA





## 4.5 Reconstruction software (ERNIE)



The MOLTEN system contains a stand-alone MS Windows program for performing the reconstructions using the training sets and numerical methods described above.

The program, ERNIE – short for “Environmental Reconstructions using Diatom Databases” – displays a list of available training sets and numerical methods. Simply choose a training set, numerical method, specify the fossil data file and output file and press Run. Note: diatom counts in the fossil data should be expressed as a percentage of the total count, and data should be in Cornell condensed or Cornell full format (see Section 4.1 above).

Optionally a translation dictionary may be supplied to convert codes in the fossil data file to MOLTEN codes. The dictionary should contain two columns of codes and be in simple text or ASCII format. The first column should contain the old code, the second column the new code. For example, the list:

AC001B AC001A

AC001C AC001A

Etc.

Converts taxon code AC001B to code AC001A etc, before applying the transfer functions.

## 4.6 Performing a TN reconstruction

All results files from ERNIE contain a header with the program name and run date / time and a summary of the input data file, training set and environmental variable, and analysis method (including verify). Note: all MOLTEN TN reconstructions are in  $\log_{10}$  ( $\mu\text{g l}^{-1}$ ) TN units. Reconstructed (hindcast) values should be back transformed to  $\mu\text{g l}^{-1}$  TN.

```
Output from program ERNIE: Environmental Reconstructions using Diatom Databases
Version 1.2: January 2004
Program run on Mon Jan 26 12:01:36 2004
```

```
Fossil data file:          C:\Data\Molten\Tool\CoresFinalPC.cep
Environmental variable:   Log_TN
Training data set:       Combined
Method:                  Verify
```



## Verify

Before performing a TN reconstruction it is recommended that the user verify the taxonomy of their sediment core diatom data against the MOLTEN training sets.

For the “Verify” method the output lists all taxa in the core file, along with their taxon number in the file (#), and the number of occurrences and maximum percentage in the core file and training set. This list should be checked for taxa that are only present in the core – indicating possible no-analog situation, or more likely, mis-coding and mis-matching between core and training set taxonomy and codes.

### Summary of taxa

Name	#	Fossil		Modern		
		No.Occ	Max	No.Occ	Max	
Ach11A	1	8	2.14	54	6.00	
AchAcaC	2	5	0.40	0	0.00	*
AchAmo	3	10	0.38	42	1.87	
AchBre	4	13	1.41	0	0.00	**
AchKri	5	1	0.34	4	3.14	
AchLacC	6	1	0.38	0	0.00	*
AchLae	7	11	0.85	7	1.39	
AchLem	8	24	0.79	37	2.05	
AchLon	9	15	0.98	0	0.00	*
AchMinuC	10	3	0.36	0	0.00	*
AchPer	11	7	0.38	0	0.00	*
AchPloI	12	1	0.51	0	0.00	*
AchRevC	13	9	0.40	43	7.53	
AchRevT	14	21	0.74	0	0.00	*
AchSpp	15	27	0.65	39	1.42	
AchSubm	16	2	0.30	7	4.19	

The Verify output also lists for each fossil sample the number of taxa (N), effective number of taxa (Hill's N2) and the sum of all taxa for each fossil sample. It also lists the number of fossil taxa at each level that are present in the training set, and the sum of taxa at each level present in the training set. In the first core sample in the example below only 22 of the 26 taxa are present in the training set, but these 22 taxa account for 98.7% of the assemblage.

### Summary of fossil samples

Sample	Data			Calibration Set	
	N	N2	Sum	N	Sum
1 H4-7000.	26	2.09	100.00	22	98.72
2 H4-7002.	28	1.62	100.00	24	98.98
3 H4-7003.	44	3.02	100.00	37	98.11
4 H4-7005.	52	4.41	100.00	41	96.92
5 H4-7007.	55	6.55	100.00	47	98.40
6 H4-7009.	35	4.45	100.00	32	99.44
7 H4-7011.	44	5.54	100.00	41	98.86
8 H4-7013.	42	5.82	100.00	36	98.00
9 H4-7015.	33	3.97	100.00	30	99.26
10 H4-7017.	38	5.43	100.00	35	98.67
11 H4-7019.	37	3.83	100.00	28	96.43
12 H4-7021.	39	2.83	100.00	34	98.48
13 H4-7023.	43	2.74	100.00	41	99.62
14 H4-7025.	44	3.12	100.00	35	97.30

**Weighted averaging partial least squares (WAPLS)**

Output for WAPLS reconstructions lists the performance statistics of the training set in terms of apparent and cross-validation (leave-one-out) errors.

Output from program ERNIE: Environmental Reconstructions using Diatom Databases  
Version 1.2: January 2004

Program run on Mon Jan 26 12:07:54 2004

Fossil data file: C:\Data\Molten\Tool\CoresFinalPC.cep  
Environmental variable: Log\_TN  
Training data set: Combined  
Method: Weighted averaging partial least squares (WAPLS)

"Apparent" errors or errors of "estimation"

Comp	RMSE	R-squared	Avg-Bias	Max-Bias
1	0.1720	0.5402	6.0691e-004	0.3458
2	0.1271	0.7489	2.9085e-004	0.4044

"Prediction" errors

Comp	RMSEP	R-squared	Avg-Bias	Max-Bias
1	0.1908	0.4361	0.0000e+000	0.4114
2	0.1599	0.6038	0.0000e+000	0.4632

The output then lists for each taxon the number of occurrences, maximum percentage, Hill's N2 and the WAPLS coefficients (beta) and their standard errors (SE) for the first 2 WAPLS components.

Species summary

		N. occur	Max	N2	Beta1	Beta2	SE1	SE2
1	Achl1A	54.0000	6.0037	26.9666	2.8903	2.9873	0.1326	0.2817
2	AchAmo	42.0000	1.8657	23.9916	3.2461	3.9744	0.1602	0.3626
3	AchJou	4.0000	1.0695	2.0720	3.0540	3.3874	0.1766	0.3287
4	AchKri	4.0000	3.1353	3.8564	2.9475	3.0588	0.2597	0.5603
5	AchLae	7.0000	1.3889	3.0151	2.7911	3.0704	0.1743	0.4844

The output then lists the reconstructed environmental values for each core level for components 1 and 2, along with the sample-specific standard errors estimated by Monte-Carlo simulation. Note that all MOLTEN TN reconstructions are in log10 ( $\mu\text{g l}^{-1}$ ) TN units and need to be back-transformed to TN  $\mu\text{g l}^{-1}$ .

Environmental reconstructions:

		Comp 1		Comp 2	
		Estimate	Std. Error	Estimate	Std. Error
1	H4-7000.	2.3263	0.2022	2.3107	0.2171
2	H4-7002.	2.3029	0.2043	2.2744	0.2271
3	H4-7003.	2.6425	0.1973	2.6754	0.1787
4	H4-7005.	2.6754	0.1948	2.6981	0.1705
5	H4-7007.	2.7331	0.1941	2.7045	0.1679



### Locally-weighted weighted-averaging (LWWA)

After the initial headers the LWWA output lists the leave-one-out cross-validation errors for the training set.

Output from program ERNIE: Environmental Reconstructions using Diatom Databases  
Version 1.2: January 2004

Program run on Mon Jan 26 12:15:35 2004

```
Fossil data file:          C:\Data\Molten\Tool\CoresFinalPC.cep
Environmental variable:   Log_TN
Training data set:       Combined
Method:                   Locally-weighted weighted averaging (LWWA)
N local:                  10
```

"Jackknife" errors or errors of "prediction"

	Inverse	Classical
RMSE	0.14129	0.13987
r2	0.69452	0.69803
Ave Bias	0.0023704	0.0032555
Max Bias	0.37838	0.35169

The LWWA output then lists the environmental reconstructions for each fossil sample, along with an estimate of the sample-specific standard error, estimated using Monte-Carlo simulation, for LWWA under inverse and classical dishrinking.

Environmental reconstructions:

		Inverse		Classical	
		Estimate	Std. Error	Estimate	Std. Error
1	H4-7000.	2.5051	0.1345	2.4784	0.1335
2	H4-7002.	2.5006	0.1405	2.4722	0.1422
3	H4-7003.	2.6354	0.1406	2.6251	0.1416
4	H4-7005.	2.6592	0.1406	2.6557	0.1416
5	H4-7007.	2.6684	0.1367	2.6624	0.1342
6	H4-7009.	2.6703	0.1506	2.6634	0.1545

### Modern analogue technique (MAT)

After the initial headers the MAT output lists the dissimilarity coefficient and number of analogues used. The default at present is the squared chi-squared distance and to take the mean or weighted mean of the 5 nearest analogues.

Output from program ERNIE: Environmental Reconstructions using Diatom Databases  
Version 1.2: January 2004

Program run on Mon Jan 26 12:20:54 2004

```
Fossil data file:          C:\Data\Molten\Tool\CoresFinalPC.cep
Environmental variable:   Log_TN
Training data set:       Combined
Method:                   Modern analog technique (MAT)
N closest:                5
```

Using Squared Chi-squared Distance measure and 5 nearest analogs

The MAT output then lists the reconstructed environmental value for each fossil sample as either the mean or weighted mean of the 5 nearest analogs, and the standard deviation (or weighted standard deviation) of the environmental value of these analogs. The output also lists the distance to the closest analog (minDC), which can be used as a measure of how good the analogs are (see section 6).



Results for Fossil data

Sample	Mean	SD	Wt-Mean	Wt-SD	minDC	N
1 H4-7000.	2.5335	0.0578	2.5479	0.0545	31.3714	5.0000
2 H4-7002.	2.5335	0.0578	2.5473	0.0555	26.1547	5.0000
3 H4-7003.	2.6180	0.0985	2.6102	0.0927	30.2809	5.0000
4 H4-7005.	2.6180	0.0985	2.6137	0.0890	37.5415	5.0000
5 H4-7007.	2.6477	0.1453	2.6553	0.1367	35.3280	5.0000
6 H4-7009.	2.6856	0.1114	2.6843	0.1067	49.6980	5.0000
7 H4-7011.	2.6726	0.1188	2.6770	0.1165	51.1387	5.0000

The MAT output also lists the 5 closest analogues for each fossil sample, their squared chi-squared distance, and their environmental value.

List of closest analogs for each fossil sample

H4-7000.	sw_H5	sw_So13	sw_B1	sw_H4	sw_No01
Distance	31.37	39.25	68.83	69.21	72.82
Log_TN	2.597	2.56	2.457	2.582	2.472
H4-7002.	sw_H5	sw_So13	sw_B1	sw_H4	sw_No01
Distance	26.15	36.6	52	56.78	67.34
Log_TN	2.597	2.56	2.457	2.582	2.472
H4-7003.	sw_H4	sw_Va04	sw_B1	fi_39	fi_61
Distance	30.28	32.24	43.37	46.53	62.48
Log_TN	2.582	2.622	2.457	2.751	2.679

## 5. References

- Birks, H. J. B. (1995). Quantitative palaeoenvironmental reconstructions. In *Statistical modelling of Quaternary Science Data*, D. Maddy and J. Brew, eds. (Cambridge: Technical Guide 5, Quaternary Research Association), pp. 161-254.
- Jones, V., and Juggins, S. (1995). The construction of a diatom-based chlorophyll a transfer function and its application at three lakes on Signy Island (maritime Antarctic) subject to differing degrees of nutrient enrichment. *Freshwater Biology* 34, 433-445.
- ter Braak, C. J. F., and Juggins, S. (1993). Weighted averaging partial least squares regression (WA-PLS): an improved method for reconstructing environmental variables from species assemblages. *Hydrobiologia* 269/270, 485-502.



## Appendix 1 List of MOLTEN dataset / transfer function contributors

Elinor Andrén Uppsala University Department of Earth Sciences & Palaeobiology Villavägen 16 SE-752 36 Uppsala, Sweden	Swedish diatom dataset
Annemarie Clarke Department of Marine Ecology National Environment Research Institute PO Box 358 4000 Roskilde, Denmark	Danish and Dutch diatom datasets
Steve Juggins School of Geography, Politics and Sociology University of Newcastle Newcastle upon Tyne NE1 7RU, UK	Transfer functions and software development
Richard Telford Bjerknes Centre for Climate Research University of Bergen N-5007 Bergen, Norway	Water chemistry database development and data screening
Kaarina Weckström Environmental Change Research Unit (ECRU) Department of Ecology and Systematics P.O. Box 65 (Viikinkaari 1) Biocentre 3, 00014 University of Helsinki Finland	Finnish diatom dataset
Ole H. Manscher National Environment Research Institute PO Box 358 4000 Roskilde, Denmark	Water chemistry data for the Danish training set
Pirkko Kauppila Finnish Environment Institute P.O. Box 140, FIN-00251 Helsinki, Finland	Water chemistry data for the Finnish training set
Department of Systems Ecology University of Stockholm SE-106 91 Stockholm Sweden	Water chemistry data for the Swedish training set
SMHI Oceanografiska Enheten Nya Varvet 31 426 71 Västra Frölunda Sweden	ditto
Analyen Nordic AB Box 11 404 404 29 Göteborg Sweden	ditto
National Institute for Coastal and Sea Research Rijkinstituut voor Kust en Zee Kortenaerkade 1 2518 Ax Den Haag The Netherlands	Water chemistry data for the Dutch training set