

Ocean Model Analysis Tools (OMAtools)

Heather J. Andres, Nancy Soontiens

Northwest Atlantic Fisheries Centre
Fisheries and Oceans Canada
80 East White Hills Road
St. John's, NL
A1C 5X1

2026

**Canadian Technical Report of
Hydrography and Ocean Sciences 424**



Fisheries and Oceans
Canada

Pêches et Océans
Canada

Canada

Canadian Technical Report of Hydrography and Ocean Sciences

Technical reports contain scientific and technical information of a type that represents a contribution to existing knowledge but which is not normally found in the primary literature. The subject matter is generally related to programs and interests of the Oceans and Science sectors of Fisheries and Oceans Canada.

Technical reports may be cited as full publications. The correct citation appears above the abstract of each report. Each report is abstracted in the data base *Aquatic Sciences and Fisheries Abstracts*.

Technical reports are produced regionally but are numbered nationally. Requests for individual reports will be filled by the issuing establishment listed on the front cover and title page.

Regional and headquarters establishments of Ocean Science and Surveys ceased publication of their various report series as of December 1981. A complete listing of these publications and the last number issued under each title are published in the *Canadian Journal of Fisheries and Aquatic Sciences*, Volume 38: Index to Publications 1981. The current series began with Report Number 1 in January 1982.

Rapport technique canadien sur l'hydrographie et les sciences océaniques

Les rapports techniques contiennent des renseignements scientifiques et techniques qui constituent une contribution aux connaissances actuelles mais que l'on ne trouve pas normalement dans les revues scientifiques. Le sujet est généralement rattaché aux programmes et intérêts des secteurs des Océans et des Sciences de Pêches et Océans Canada.

Les rapports techniques peuvent être cités comme des publications à part entière. Le titre exact figure au-dessus du résumé de chaque rapport. Les rapports techniques sont résumés dans la base de données *Résumés des sciences aquatiques et halieutiques*.

Les rapports techniques sont produits à l'échelon régional, mais numérotés à l'échelon national. Les demandes de rapports seront satisfaites par l'établissement auteur dont le nom figure sur la couverture et la page de titre.

Les établissements de l'ancien secteur des Sciences et Levés océaniques dans les régions et à l'administration centrale ont cessé de publier leurs diverses séries de rapports en décembre 1981. Vous trouverez dans l'index des publications du volume 38 du *Journal canadien des sciences halieutiques et aquatiques*, la liste de ces publications ainsi que le dernier numéro paru dans chaque catégorie. La nouvelle série a commencé avec la publication du rapport numéro 1 en janvier 1982.

Canadian Technical Report of
Hydrography and Ocean Sciences 424

2026

Ocean Model Analysis Tools (OMAtools)

by

Heather J. Andres, Nancy Soontiens

Northwest Atlantic Fisheries Centre
Fisheries and Oceans Canada
80 East White Hills Road
St. John's, NL
A1C 5X1

© His Majesty the King in Right of Canada, as represented by the Minister of the Department of Fisheries and Oceans, 2026.

This work is licensed under the [Open Government Licence](#)

Cat. No. Fs 97-18/424E-PDF ISBN 978-0-662-79566-7 ISSN 1488-5417

<https://doi.org/10.60825/me4y-zd67>

Correct citation for this publication:

Andres, H. J., Soontiens, N. 2026. *Ocean Model Analysis Tools (OMAtools)*. Can. Tech. Rep. Hydrogr. Ocean Sci. 424: vii + 29 p. <https://doi.org/10.60825/me4y-zd67>

Contents

List of Tables	iv
List of Figures	v
Abstract	vi
Résumé	vii
1 Introduction	1
2 Methods	2
2.1 Input data requirements and interpolation	4
2.2 Comparisons over a single time step	5
2.3 Time-dependent comparisons	6
2.3.1 Climatological bias	7
2.3.2 Variance comparisons	8
2.3.3 Correlation	10
2.3.4 Gamma-squared (aggregate skill metric)	10
3 Data	11
3.1 Canadian Atlantic Shelf Temperature-Salinity dataset	11
3.2 GLORYS12v1	12
4 Examples and Applications	13
4.1 Example 1: comparing Stn27 temperature and salinity profiles	13
4.2 Example 2: comparing measurements and model on the Seal Island transect	18
4.3 Example 3: comparing GLORYS12v1 user-grid bathymetry against the depth of the deepest temperature	23
4.4 Example 4: comparing different representations of modelled bottom temper- atures on the NL Shelf	24
5 Conclusions	25
6 Code Availability	28
7 Acknowledgments	28

List of Tables

1	Time-invariant metrics related to comparisons at Stn27 in Example 1. CASTS includes 90 profiles within the years 1993-2024 during the winter, 392 profiles during the spring, 216 profiles during the summer, and 347 profiles for temperature and 348 for salinity during the fall.	17
2	Time-invariant metrics related to comparisons at Seal Island in Example 2. There are 31 transect observations of temperature and salinity during fall and 69 temperature transects and 67 salinity transects during summer.	20
3	Time-invariant metrics related to bathymetry differences in Example 3. . . .	23
4	Time-invariant metrics related to bottomT and deepestT comparisons in Example 4.	24

List of Figures

1	Flow chart representing operation of OMAtools functions.	3
2	Time-averaged Stn27 temperature (left) and salinity (right) profiles for AZMP in dots and GLORYS12v1 in x's. A vertical black line in the temperature plot indicates 0°C. Seasons are indicated by dot or x colour, where winter (grey) corresponds to January-February-March, spring (blue) corresponds to April-May-June, summer (green) corresponds to July-August-September, and fall (orange) corresponds to October-November-December.	15
3	Metrics for temperature (left) and salinity (right), where filled dots are statistically significantly different than the reference, and x's are statistically insignificant values. The metrics include climatological bias (GLORYS12v1-AZMP, top row) as a function of depth at Stn27, with vertical reference lines at values of 0 (units dependent on subplot); temporal correlation (second row) with a reference value of 0; normalized variance difference (GLORYS12v1-AZMP, third row) with vertical reference lines at values of 0; and gamma-squared (bottom row) with vertical reference lines at values of 1. Colours indicate season as shown in the legend.	16
4	Temperature (left) and salinity (right) climatologies on the SI transect for summer (top row) and fall (bottom row). CASTS data is plotted in colours, and GLORYS12v1 data is plotted in black contours for the same levels. Dashed contours indicate negative values.	19
5	Temperature (left) and salinity (right) metrics for the SI transect for the summer (July-August-September). The metric is labelled on each plot, and grey hatching indicates values insignificantly different than 0 for climatological bias (units either °C or g/kg), temporal correlation and normalized variance difference (both unitless). Insignificant values for gamma-squared are relative to a value of 1.	21
6	Temperature (left) and salinity (right) metrics for the SI transect for fall (October-November-December). The metric is labelled on each plot, and grey hatching indicates values insignificantly different than 0 for climatological bias (units either °C or g/kg), temporal correlation and normalized variance difference (both unitless). Insignificant values for gamma-squared are relative to a value of 1.	22
7	Static bathymetry field on the user grid (left), the depth of the deepest user grid model level with valid data (middle), and the difference between the two (right).	24
8	Climatologies of bottom temperature on the C-grid (left), deepest temperature based on the user-grid three-dimensional temperature field (centre), and bottom temperature on the user grid (right) from GLORYS12v1 for all days from 1993 to the end of 2019.	25
9	(Top left) Climatological bias for “deepestT” minus “bottomT”, (top right) correlation with time, (bottom left) normalized variance difference, and (bottom right) gamma-squared for “bottomT” and “deepestT” on the user grid.	26

ABSTRACT

Andres, H. J., Soontiens, N. 2026. *Ocean Model Analysis Tools (OMAtools)*. Can. Tech. Rep. Hydrogr. Ocean Sci. 424: vii + 29 p. <https://doi.org/10.60825/me4y-zd67>

The Ocean Model Analysis tools (OMAtools) python package is designed to provide quick comparisons of modelled and observed oceanographic fields. It includes functions that cover a range of data attributes (climatological bias, temporal and spatial correlation, difference in variance, and γ^2), and it provides estimates of the significance of the results. It is designed for a range of input geometries from 1-dimensional profiles to 3-dimensional fields. The package is not intended to provide an exhaustive comparison of the two input datasets, but its goal is to provide a broad evaluation on a range of characteristics that could then be used to initiate more tailored analyses.

This report defines the metrics available in OMAtools and how significance is assessed. It outlines the requirements and limitations of the tools and provides examples of how they can be applied in the Newfoundland and Labrador region of Fisheries and Oceans Canada using data from the GLORYS12v1 ocean reanalysis and the Canadian Atlantic Shelf Temperature-Salinity dataset.

RÉSUMÉ

Andres, H. J., Soontiens, N. 2026. *Ocean Model Analysis Tools (OMAtools)*. Can. Tech. Rep. Hydrogr. Ocean Sci. 424: vii + 29 p. <https://doi.org/10.60825/me4y-zd67>

Le progiciel Python Ocean Model Analysis tools (OMAtools) est conçu pour fournir rapidement comparaisons des champs océanographiques modélisés et observés. Il comprend des fonctions qui couvrent une gamme d'attributs de données (biais climatologique, corrélation temporelle et spatiale, différence de variance et γ^2), et il fournit des estimations de l'importance des résultats. Il est conçu pour une gamme de géométries d'entrée allant des profils à 1 dimension aux champs à 3 dimensions. Le ne vise pas à fournir une comparaison exhaustive des deux ensembles de données d'entrée, mais son objectif est de fournir une évaluation générale d'un éventail de caractéristiques qui pourraient ensuite être utilisées pour lancer des analyses plus personnalisées.

Ce rapport définit les indicateurs disponibles dans OMAtools et la manière dont la signification est évaluée. Il décrit les exigences et les limites des outils et fournit des exemples de la façon dont ils peuvent être appliqué dans la région de Terre-Neuve-et-Labrador de Pêches et Océans Canada en utilisant données de la réanalyse océanique GLORYS12v1 et de l'ensemble de données Canadian Atlantic Shelf Temperature-Salinity.

1 Introduction

The Ocean Model Analysis tools (OMAtools) is a python package designed for ocean model-data comparisons. OMAtools includes functions that span data features that users commonly wish to compare (e.g. biases and temporal correlations) and provides assessments of the statistical significance alongside results, while being general enough to apply to multiple different variables and geometries. OMAtools functions are designed to provide a quick glance at the differences between two datasets as a starting point for a user’s more tailored analyses. OMAtools has dependencies on the xarray python package (among others) and uses many of its functions. In the examples presented here, other packages like the Gibbs SeaWater (GSW) Oceanographic Toolbox (McDougall, Barker 2011) are used effectively in conjunction with OMAtools.

Metrics for time-varying datasets provided by OMAtools include the climatological bias, temporal correlation, normalized variance difference, and γ^2 . While the first two metrics are widely used, the latter are less so. One example of normalized variance difference assessments occurring is found in Taylor diagrams (e.g. Castillo-Trujillo et al. 2023, compare the magnitude of the temporal variance between observations and model results, when both are normalized to the variance of the observations). γ^2 (or gamma-squared) is a commonly used summary metric among modellers. It is the ratio of the variance of the difference between the modelled and observed data divided by the variance of the observations (Thompson et al. 2003), and it effectively compiles variance magnitude and correlation comparisons into a single metric. Note that it is not very useful in cases when the variance of the observations is near zero, due to that term being in the denominator. Together, this set of metrics provides insight into the accuracy or quality of the system relative to observations (Ryan et al. 2015). It also addresses key questions commonly explored in model evaluations:

- is the mean condition the same as observed,
- does the modelled variable reproduce the timing and sign of observed changes,
- are the magnitudes of the modelled variations consistent with those observed.

A metric on its own without an assessment of the significance of the result is of limited usefulness. For example, a correlation value of 0.5 might indicate an important relationship between two variables, but it may reasonably arise by chance if the correlation is calculated over only five observations. OMAtools functions assess the significance of metric results using the statistical characteristics of the datasets provided to the metrics functions. Using the sample variances, covariances, and numbers of degrees of freedom combined with standard statistical tests, significance estimates are returned with metric results. Note that the OMAtools metrics may be applied to pairs of 1D, 2D, or 3D fields, producing maps of differences (as in Griffies et al. 2009), as shown in Figure 1. Although the statistical tests are applied independently in each grid cell, treating the results of the tests as independent may not accurately reflect the inter-relatedness of oceanographic conditions in nearby grid cells. One approach to interpreting the significance of the field of results that considers the multi-dimensionality of the datasets and tests, known as the Bonferroni correction, is described in Section 2.3. However, it is left up to the user to interpret the meaning of the significance of

the results provided and consider how these questions apply to their own data. Additionally, each test involves assumptions about the properties and suitability of the data they are applied to. These assumptions are detailed in this report and in OMAtools docstrings.

Finally, the user is reminded that model-data comparisons require bridging the inherent differences in these two types of data in order to understand whether the representations are “consistent enough” for the user’s purpose. We present as an example temperature profiles from an ocean general circulation model and from conductivity, temperature, depth (CTD) measurements. CTD casts taken on the continental Newfoundland and Labrador Shelf as part of the Atlantic Zone Monitoring Program (AZMP, see Section 3.1 for more information) typically take on the order of 5 to 10 minutes to complete with measurements every 5 m using sensors with sizes on the order of centimetres. These spatially and temporally localized measurements are contrasted against a modelled profile such as from GLORYS12v1 (see Section 3.2 for more details), which is based on conditions evaluated over grid cells with typical horizontal extents of approximately 6 km in this region and thickness between 0.5 m and 100 m. The model data is representative of conditions over a time step, and for GLORYS12v1, is further averaged into daily timescales. The spatial and temporal scales of grid cells can alter the nature of signals that can be represented (resolved versus unresolved phenomena, often with parametrizations to compensate for missing sub-grid processes), but the models present a consistent representation in time and space. In contrast, the observations provide a snapshot of a complete physical system, but with gaps in space and time. Finally, the variables may represent different phenomena and not be entirely comparable (e.g. McDougall, Barker, et al. 2021), especially when observations do not directly measure the variable itself but a proxy from which a variable can be reconstructed.¹ None of the uncertainties associated with these types of data differences are addressed by OMAtools functions.

In summary, OMAtools provides a set of functions that evaluate commonly used metrics when running model-data comparisons. They are designed to be generally applicable to a wide range of variables and geometries, and metrics are returned with estimates of the statistical significance of the results. Details about the methods used in the metric functions can be found in Section 2. A range of examples illustrating the use of these functions are provided in Section 4 using datasets described in Section 3.

2 Methods

OMAtools functions are organized according to the spatial and temporal attributes of the comparison data. Temporal metrics differ depending on whether they are evaluated on a single time instance (e.g. climatologies) or time-varying data. For each of these temporal cases, functions are generalized to permit multiple options for the spatial characteristics of the input data, as long as the two datasets being compared have the same dimensionality and include at least one spatial dimension. Outputs from time-varying functions are formatted as `xarray.Datasets`, where different metrics are provided as variables. Outputs from the time-invariant functions are formatted as lists. In neither case is metadata retained from

¹This issue is particularly evident in paleoclimate studies where the connection between the measurement and the variable of interest may be more indirect. See Weitzel et al. (2024) for an example of how such uncertainty is addressed in one context.

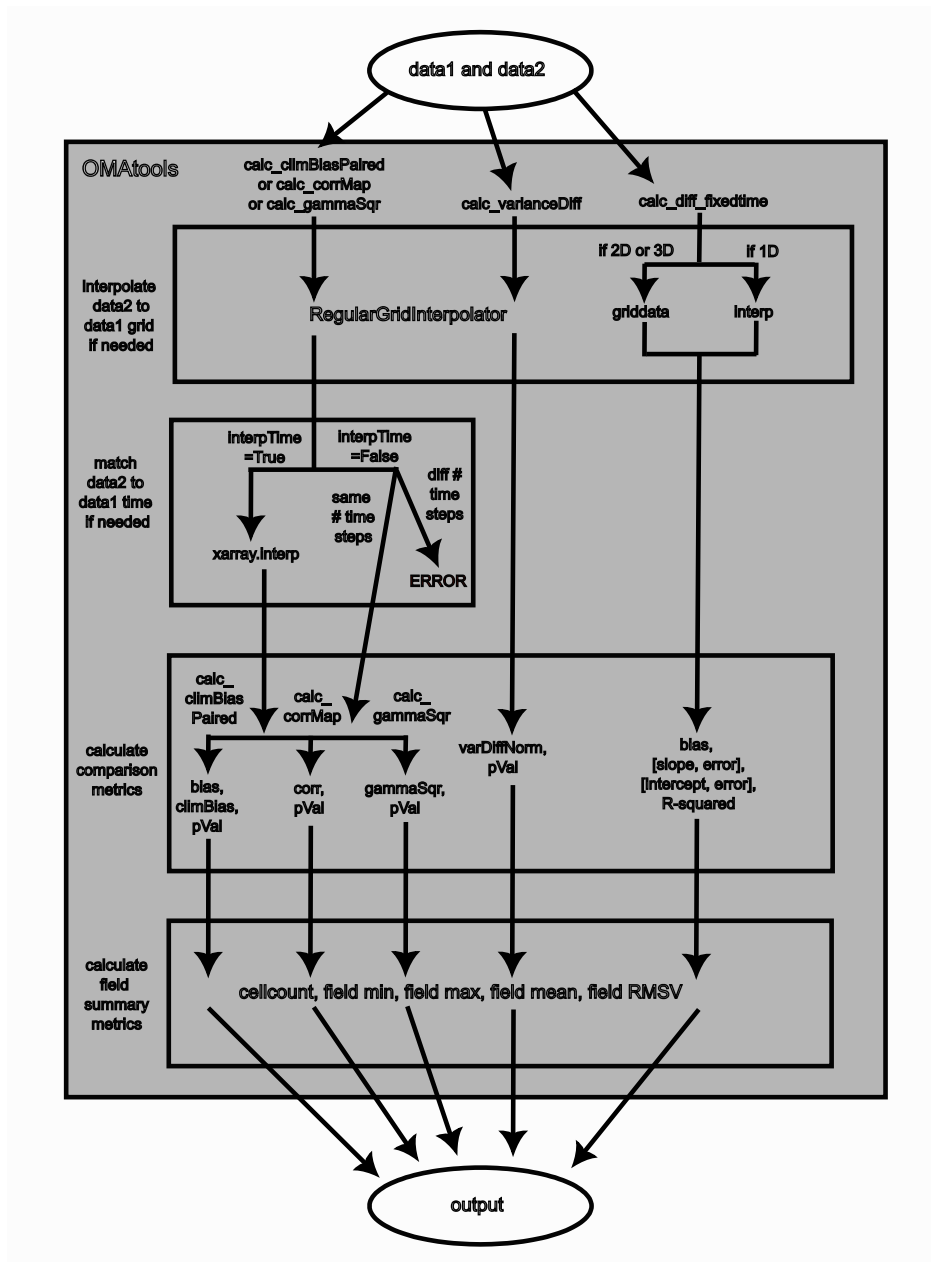


Figure 1: Flow chart representing operation of OMAtools functions.

the input data, and no metadata is included regarding the analysis performed by OMAtools. Including informative metadata in OMAtools outputs has been identified as a task for future work.

The general flow of OMAtools functions are illustrated in Figure 1 and described in detail below.

2.1 Input data requirements and interpolation

OMAtools requires that two input datasets are provided to comparison functions, with the convention that the first dataset is treated as a reference dataset (like observations), and the second represents a comparison dataset (such as might be obtained from model simulations). The two datasets must have the same type. All OMAtools functions accept data formatted as `xarray.DataArrays`, but the fixed-time functions also accept some other data formats detailed in their descriptions below. Both datasets must have the same dimensionality and dimension names (e.g. two-dimensional data with dimensions of "latitude" and "longitude"), but they do not need to have the same grid. Acceptable options for the dimensionality of the input data includes one (e.g. vertical profiles), two (e.g. sea surface temperature fields), or three dimensions (e.g. depth-dependent salinity fields). If the two datasets have matching grids, then there are no additional input data requirements. Indeed, although OMAtools are designed for model-observational data comparisons, there is nothing preventing a user from comparing two observational datasets or two model datasets that meet these criteria.

If the two input datasets do not have matching spatial grids, then additional input data formatting restrictions apply. Most of the OMAtools functions compare two datasets grid cell to grid cell, so the two datasets need to be on the same grid. As a result, **OMAtools provides spatial interpolation functions to interpolate the second (comparison) dataset to the grid of the first (reference) dataset**. These built-in interpolation tools permit users to perform as little preparation of the data as they would like. However, some grid configurations, like curvilinear or irregular grids, are not handled by the interpolation schemes used within OMAtools comparison functions. We flag to users that OMAtools does not fail gracefully in this circumstance. The reason is that OMAtools currently only accepts one-dimensional coordinate variables to describe grid cells and not multi-dimensional dimensions like curvilinear latitudes or longitudes. Thus, OMAtools does not recognize curvilinear grids and will by default attempt interpolation on the non-physical coordinates if they differ. That is why users must consider data preparation choices carefully and ensure curvilinear datasets are already on the same grid prior to calling OMAtools.

For input datasets in `xarray.DataArray` format, the spatial interpolation is performed by the function `regrid_via_RGI` using SciPy's `RegularGridInterpolator` function (The SciPy community 2008b). It requires that input data be on a rectilinear grid with grid-cell edges aligned along latitudinal and longitudinal directions but not necessarily uniformly spaced. In addition, spatial coordinates must be either monotonically increasing or decreasing. This method does not perform well when the scales of one dimension's increments are equivalent to or larger than the range of another dimension, so dimension units may need to be scaled prior to calling OMAtools functions if interpolation will be required. Finally, a current limitation associated with the use of `RegularGridInterpolator` is that it automatically interpolates over all dimensions simultaneously. **Thus, if a user has provided datasets with different**

time dimensions that also require spatial interpolation, temporal interpolation will be performed automatically, regardless of whether the user requested it or not. If a user wishes to avoid this behaviour, the user would be advised to spatially interpolate the datasets offline prior to calling OMAtools. When all of these requirements are satisfied, then the data is multi-dimensionally interpolated to the new grid points using a linear approach by default, although the user may specify a different method. For the examples presented in this report, a nearest neighbour approach was selected to reduce the loss of data coverage near coastlines and the ocean bottom. RegularGridInterpolator’s multi-dimensional linear interpolation scheme sets any interpolated values to NaN that have a single NaN value within its multi-dimensional cloud of surrounding points, even if the input file contains valid data at the interpolation location. For points near the bottom or near a land-ocean boundary, this can result in a loss of data. In contrast, nearest neighbour interpolation only sets extrapolated points and any points with a NaN as one of the nearest neighbours to NaN.

For spatially interpolating data that are not formatted as `xarray.DataArray` (accepted formats include list and `numpy.ndarray`), the function *ND_remap* is used. *ND_remap* performs a linear remapping by default using NumPy’s `interp` function (NumPy Developers 2008-2025) for one-dimensional data and SciPy’s `griddata` function (The SciPy community 2008a) for all other cases. NumPy’s `interp` function requires the spatial coordinate variable to be increasing and free of NaNs. SciPy’s `griddata` function is designed to work with unstructured data, so it has no additional input data formatting requirements beyond those of OMAtools. However, it is slower than RegularGridInterpolator when applied to structured data. Other interpolation options are available for *ND_remap* when `griddata` is used, including “nearest” or “cubic”.

In addition to spatial interpolation, some metrics require that the time be matched between the two input datasets (e.g. the climatological bias, correlation and gamma-squared). Within OMAtools, matched times may either mean having the same time values in both datasets or having the same number of timesteps with temporal interpolation turned off. One example of when the latter case may be selected is if a user wanted to compare two different years’ worth of daily temperatures: although the particular time values differ, the number of time steps is the same, and temporal interpolation would not be wanted. When neither of these matching conditions are met, temporal interpolation must be used for the functions requiring matched times to run successfully. Temporal interpolation is provided in function *testCoords_NDtime* using Xarray’s linear interpolation function.

2.2 Comparisons over a single time step

Time-independent metrics are applied to input datasets with no time dimension. Two functions control the evaluation of time-independent metrics: *calc_metrics_fixedtime* and *calc_diff_fixedtime*.

For situations where a user already has a single 1D, 2D, or 3D field of comparison metric results, *calc_metrics_fixedtime* generates a set of field summary metrics that include the number of grid cells with valid data, field minimum, field maximum, field mean and field root-mean-square value (RMSV). A sample call to *calc_metrics_fixedtime* is

```
output = calc_metrics_fixedtime(data,verbose)
```

where output is a list containing the metric values in the order listed above, data is an input dataset formatted as either a numpy.ndarray, or a pandas.DataFrame, or an xarray.DataArray. Verbose is a boolean flag indicating whether additional debugging output is desired to be printed to the standard out.

When a user wants to evaluate the differences between two time-invariant datasets (1D, 2D, or 3D) that may not be on the same spatial grid, *calc_diff_fixedtime* can be used. It spatially interpolates the comparison dataset onto the reference dataset grid as needed, calculates their differences in every grid cell and then calls *calc_metrics_fixedtime* to generate summary metrics for the field. In addition, this function performs an ordinary least-squares, linear, spatial regression of one input dataset against the other and returns the regression slope and its error, the regression intercept and its error, and the percentage of explained variance (or R^2) as additional output metrics. The field of differences is also returned. A sample call to *calc_diff_fixedtime* is as follows.

```
outmetrics,outdiff = calc_diff_fixedtime(data1,data2,x1,x2,y1,y2,z1,
                                         z2,interpype="nearest",verbose=True)
```

In this case, data1 is the reference dataset, data2 is the comparison dataset, and outdiff is their difference with a sign corresponding to data2 minus data1. Data1 and data2 may be either lists, numpy.ndarrays or xr.DataArrays, but they must be the same type, and outdiff is returned in the same format. The coordinates of the data spatial axes are provided by x, y and z, where x is the fastest-varying dimension, y the next fastest, and z the slowest (i.e. data1(z1,y1,x1)). In this example, all three coordinates are provided for the two datasets, since data1 and data2 are three-dimensional in space. However, if the datasets had fewer dimensions, then the z's (corresponding to the two-dimensional case of data1(y1,x1)) or z's and y's (for the 1-dimensional case, data1(x1)) would not be provided. Interptype is an optional argument indicating what type of interpolation is preferred if it is needed. Its default value is "linear", so it need only be provided if the user wishes to select a different option. Finally, verbose is an optional argument, which results in additional debugging output being provided when set to True. For this function, outmetrics is a list that includes all of the output metrics of *calc_metrics_fixedtime*, followed by a two-element list for the ordinary least-squares linear regression slope for data2 on data1 and its error, a two-element list of the regression intercept and its error, and the R^2 .

2.3 Time-dependent comparisons

Time-dependent functions require input datasets to be formatted as xarray.DataArrays with a dimension named "time". Functions that evaluate time-varying data include *calc_climBiasPaired*, *calc_varianceDiff*, *calc_corrMap*, and *calc_gammaSqr*.

Note that all time-varying metrics are calculated independently in each grid cell, and statistical tests are applied as if conditions in each grid cell varied independently of neighbouring locations. This is likely a poor assumption in many oceanographic contexts and can

lead to erroneous identifications of significance. For example, in a field of 100 grid cells with random data, it is expected that a t-test applied at 95% significance separately in each grid cell will lead to five grid cell values identified as significant by chance. One way of minimizing the impact of this assumption is to apply a Bonferroni correction (Johnson, Wichern 2007) to the significance assessments. This correction decreases the p-value threshold from 0.05 (for 95% significance) in each grid cell to 0.05/M where M is the number of independent grid cells with data in the field. In this case, a grid-cell p-value would be identified as significant if its value was less than 0.05/M. For the functions presented here, p-values for each metric in each grid cell are returned with the metric, and it is left up to the user to choose how to interpret the results and whether they wish to apply a Bonferroni or some other correction.

2.3.1 Climatological bias

calc_climBiasPaired calculates the climatological bias in every grid cell and its significance.

```
output = calc_climBiasPaired(data1,data2,interpTime=False,
                             interptype="nearest",verbose=True)
```

Provided two input time-varying datasets *data1* and *data2* with spatial dimensions ranging from 1D to 3D and formatted as *xarray.DataArrays*, the function spatially interpolates the comparison dataset to the grid of the reference dataset if necessary using the interpolation method specified by *interptype*. “*interpTime*” is an optional argument with a default value of *False* that indicates whether the user wants temporal interpolation to be used for time matching or not. “*interptype*” is an optional parameter with a default value of “*linear*”, so it is only specified in the example function call to change this method to nearest neighbour. *calc_climBiasPaired* requires that the time coordinates match between *data1* and *data2* as described in Section 2.1. A field of differences for each time step is then calculated, and an average over all time values is produced. This average represents the climatological bias in each grid cell, which is returned by the function. The output of a call to *calc_metrics.fixedtime* for the field of climatological biases is also returned. These outputs are returned in a single *xarray.Dataset* with dimensions matching those of *data1*, as well as a new dimension named “*values*” with coordinate values of “*value*” and “*error*”. The variables in “*output*” include:

- “*bias*”, which has the same dimensions as *data1* and represents the time-varying differences between *data1* and *data2* (sign corresponding to *data2* minus *data1*),
- “*climBias*”, which has the same spatial dimensions as *data1* and is the time-averaged difference between *data1* and *data2*,
- “*pVal*”, which has the same dimensions as *climBias* and represents the likelihood that the *climBias* value in each cell is consistent with a null hypothesis of 0 (see below),
- “*cellcount*”, which is a function of “*values*”, with the number of grid cells with valid data in “*climBias*” in the first coordinate entry,

- “fieldmin”, which has dimension “values” with the minimum value of “climBias” in the first coordinate entry,
- “fieldmax”, which has dimension “values” with the maximum value of “climBias” in the first coordinate entry,
- “fieldmean”, which has dimension “values” with the mean value of “climBias” in the first coordinate entry,
- “fieldRMSE”, which has dimension “values” with the root-mean-squared error of “climBias” in the first coordinate entry.

The likelihood of a particular climatological bias being obtained when its underlying distribution has an expected value of zero is represented by “pVal”. It is assessed via the paired t-test, whose statistic is:

$$T(\vec{X}) = \frac{\overline{\zeta_{diff}(\vec{X})}}{\frac{\sigma_{diff}(\vec{X})}{\sqrt{N(\vec{X})}}}. \quad (1)$$

In Equation 1, $\overline{\zeta_{diff}(\vec{X})}$ is the difference between the two datasets averaged over all $N(\vec{X})$ valid time steps in the grid cell given by $\vec{X}$, and $\sigma_{diff}(\vec{X})$ is the standard deviation of the differences in that grid cell. The degrees of freedom associated with the t-statistic are $N(\vec{X}) - 1$. If there are not at least three temporal values in a grid cell, then the significance for that grid cell is set to NaN. Everywhere else, double-sided p-values are calculated from the t-distribution using the t-statistic and degrees of freedom. An array of p-values is returned in the shape of the spatial field of the first dataset.

Note that paired t-tests apply when the observations in one sample are understood to be matched in some way to those in the second sample. Examples can include model output and observations for the same date in time, or observations for the same day of year from two different years, or two sets of model output for the same date with modified forcings. Additional requirements are that the samples are random and that the differences between the paired data are approximately Gaussian and free of outliers.

2.3.2 Variance comparisons

In addition to tests of the first moment of the data, the mean, comparisons are made of the second moment, the variance. As with the bias calculation, two input time-varying datasets with spatial dimensions ranging from 1D to 3D are passed to *calc_varianceDiff*. If the spatial coordinates do not match between the two datasets, the second dataset is remapped to the grid of the first. However, there is no requirement that the datasets match in time as was required for the paired bias comparisons, so temporal interpolation is not used. Nor do the numbers of data points in the two datasets need to be equal. Instead, there is a requirement that at least three time values be provided for each dataset.

Variances are calculated independently per grid cell for each dataset, and both variances are scaled by the best estimate of the variance in the grid cell. The reason for scaling the variances is that interpreting variance differences can be difficult without knowing what

typical variance values are. When the variances are scaled, then differences can be easily understood in terms of the best estimate of the variance of the grid cell. For example, a scaled variance difference of 2 would mean the difference between the variances of the two datasets is twice the best estimate of the variance of that grid cell. There are two choices of which variance to use to scale the data. When a user considers the variance of the first dataset to be more representative of the system, such as if the first dataset contains observed data and the second modelled data, they can choose to scale both datasets by the variance of the first dataset. Alternatively, when both datasets are thought to provide equally valid representations of the variance of the system, a user can choose to scale both datasets by the average of the two variances. Once the scaling is complete, the difference in the variances is calculated.

A sample call is as follows.

```
output = calc_varianceDiff(data1,data2,normcase,
                           interptype="nearest",verbose=True)
```

As previously, *data1* and *data2* are the reference and comparison datasets, respectively. For *calc_varianceDiff*, they must both be *xarray.DataArrays* with the same dimensions, although their coordinate values do not need to match. In addition to *interptype* and *verbose*, which have the same representations as in *calc_climBiasPaired*, there is an additional argument "normcase" indicating the normalization option. An integer value of 1 indicates that the variance of *data1* is used to scale the variance difference, while an integer value of 2 indicates that the average of the variances of *data1* and *data2* will be used to scale the variance difference. The output *xarray.Dataset* has the same structure as for *calc_climBiasPaired*, except that instead of "bias" and "climBias", a variable named "varDiffNorm" is returned, and "fieldRMSE" is replaced by "fieldRMSV" to represent the root-mean-squared value.

The significance term for this function indicates the likelihood that the variance difference is not zero. This assessment is performed via the Levene homogeneity test, which evaluates the likelihood that variances from *k* samples, each with N_i measurements, are the same (Filliben, Heckert 2012). It is applied here for the specific case of two samples via the following equation:

$$W_{Levene}(\vec{X}) = (N - 2) \frac{\sum_{i=1}^2 [N_i (\overline{Z_i(\vec{X})} - \langle Z(\vec{X}) \rangle)^2]}{\sum_{i=1}^2 \sum_{j=1}^{N_i} (Z_{i,j}(\vec{X}) - \overline{Z_i(\vec{X})})^2}, \quad (2)$$

where $Z_{i,j}(\vec{X})$ is given by

$$Z_{i,j}(\vec{X}) = |\zeta_{i,j}(\vec{X}) - \overline{\zeta_i(\vec{X})}| \quad (3)$$

and $\zeta_{i,j}$ is the *j*th scaled variance value in the grid cell given by $\vec{X}$ for dataset *i*. *N* represents the sum over all N_i . Note that the overbar in Equation 3 represents either a mean over time, which is appropriate for symmetric distributions with moderately sized tails, or a median over time, which is better for skewed distributions and is generally recommended. In OMAtools, the median is used. In Equation 2, $\overline{Z_i(\vec{X})}$ represents a mean over time, while $\langle Z(\vec{X}) \rangle$ represents a mean over time and sample. The significance of the Levene statistic is evaluated against the F distribution with 1 and $N - 2$ degrees of freedom.

2.3.3 Correlation

The third metric function is *calc_corrMap*, which calculates the temporal correlations between the two datasets in each grid cell.

```
output = calc_corrMap(data1,data2,interpTime=False,
                      interptype="linear",verbose=False)
```

As for the other time-varying metric functions, *data1* and *data2* must be formatted as *xarray.DataArrays*, and the second dataset is interpolated to the spatial grid of the first if necessary. Correlations require that the two datasets have matched times as defined in Section 2.1. Whether the time matching is achieved via interpolation or via an equal number but different time steps is indicated by the third field, “*interpTime*”: True indicates that interpolation is requested, while False indicates that interpolation is not wanted. “*interptype*” and “*verbose*” are as described previously. The output *xarray.Dataset* is structured as for *calc_varianceDiff*, except that instead of the variable “*varDiffNorm*” being returned, the variable is named “*corr*”.

Once the two datasets are on corresponding spatial and temporal grids, the correlations between the two datasets are calculated in each grid cell where there are at least 3 time steps. The likelihood that the correlation between the underlying distributions is zero given a particular correlation result is assessed via the t-statistic for correlations:

$$T(\vec{X}) = r(\vec{X}) \sqrt{\frac{N(\vec{X}) - 2}{1 - r(\vec{X})^2}}. \quad (4)$$

In Equation 4, $r(\vec{X})$ is the correlation and $N(\vec{X})$ is the number of temporal data points common to the two datasets in grid cell $\vec{X}$. The t-statistic is evaluated relative to the t-distribution for $N(\vec{X})-2$ degrees of freedom to generate the p-value, which is returned from the function.

2.3.4 Gamma-squared (aggregate skill metric)

The final metric function applied to time-varying fields is *calc_gammaSqr*, which evaluates gamma-squared independently in each grid cell. Gamma-squared in grid cell $\vec{X}$ is defined by Equation 5 (Thompson et al. 2003).

$$\gamma^2(\vec{X}) = \frac{\sigma_{diff}^2(\vec{X})}{\sigma_{\zeta_1}^2(\vec{X})} \quad (5)$$

$\sigma_{diff}^2(\vec{X})$ is the variance of the difference between the comparison data (ζ_2) and the reference data (ζ_1) in grid cell $\vec{X}$, and $\sigma_{\zeta_1}^2(\vec{X})$ is the variance of the reference data there. A perfect match between comparison and reference data would yield a gamma-squared value of zero, while a constant estimator replacing the modelled data would yield a value of one. Thus, values close to zero indicate good model performance, and values larger than one indicate particularly poor model performance.

A sample function call is as follows.

```
output = calc_gammaSqr(data1,data2,interpTime=False,
                       interptype="linear",verbose=False)
```

In this example, `data1` and `data2` are the reference and comparison datasets, respectively, formatted as `xarray.DataArrays`. As for previous metric calculations, the second dataset is interpolated to the spatial grid of the first if necessary. Since calculating gamma-squared requires evaluating the variance of differences during at least three timesteps, the two datasets must be matched in time as described in Section 2.1. “`interpTime`”, “`interptype`” and “`verbose`” are optional arguments described previously. The output of *calc_gammaSqr* is an `xarray.Dataset` formatted identically to the output of *calc_corrMap*, except that variable “`corr`” is replaced by variable “`gammaSqr`”.

The significance of the gamma-squared result is assessed independently in every grid cell through a comparison against the F-distribution with $N-1$ degrees of freedom for both numerator and denominator. The null hypothesis is that gamma-squared equals a value of 1. Two alternative hypotheses are proposed: that the gamma-squared is larger than 1 (i.e. significantly poorer performance than a constant estimator), or that it is less than 1 (i.e. significantly better performance than a constant estimator). The significance value represents the likelihood that the null hypothesis is true and that the given gamma-squared value was obtained by chance. A sign convention is introduced to indicate which one-tailed test the returned significance values refer to. If the significance refers to a test of gamma-squared values greater than 1, it is positive. If the significance refers to a test of gamma-squared less than 1, it is made negative.

3 Data

Two different datasets are used for all of the application examples presented. The first contains observational profiles obtained from the Canadian Atlantic Shelf Temperature-Salinity (CASTS) dataset. The second is gridded model data from the GLORYS12v1 oceanographic reanalysis product. These two datasets are described below.

3.1 Canadian Atlantic Shelf Temperature-Salinity dataset

The CASTS dataset is described in Coyne, Cyr, Atchison, et al. (2026). It compiles over eight hundred thousand oceanographic temperature and salinity profiles in the North-West Atlantic and Eastern Arctic (between 35°N to 80°N, 42°W to 100°W) that are sourced from twelve existing databases. The data from these different sources were processed to comparable levels of quality control and averaged onto a standard set of pressure levels (1 dbar spacing from the surface to 1000 dbar, 10 dbar spacing from 1000 dbar to 2000 dbar, and 100 dbar spacing from 2000 dbar to 5000 dbar). Since not all profiles included measurements at all of these levels, many profiles are filled with NaNs at the locations of these data gaps in the vertical. The ensuing profiles were then batched according to their year of data collection and written to netCDF files, where the time of the profile and the vertical level are the only dimensions. The latitudinal and longitudinal positions of the profile, as well as temperature and salinity measurements and other attributes are variables in these annual

files. Finally, additional quality checks were performed on the annual files before they were archived (Coyne, Cyr, Donnet, et al. 2025).

Most of the examples presented here involve analyses on Atlantic Zone Monitoring Program (AZMP) transects or stations in the Newfoundland and Labrador (NL) region obtained from the CASTS dataset. The AZMP is an oceanographic observational programme of Fisheries and Oceans Canada (DFO) that began in the late 1990s and continues to the present (Therriault et al. 1998). In the Newfoundland and Labrador Region of DFO, the AZMP monitors conditions at nine transects up to three times a year and more frequently at Station 27 (Stn27, 47°32.8'N and 52°35.2'W). Data from Stn27 and the Seal Island (SI) cross-shelf transect located in southern Labrador are analysed in Examples 1 and 2 in Section 4.

For both station and transect data, the same processing steps are used for CASTS data prior to calling OMAtools. All CASTS data within 0.8 km of station coordinates were identified as being located at the station. Only this data was kept for further analysis. Next, the data was gridded to meet the requirements specified in Section 2.1. If only a single station were being examined, the data was re-organized to be time by depth by latitude by longitude, where latitude and longitude coordinates had lengths of 1. This step is unnecessary if the model data is already located at the same station and no interpolation is required. However, in Section 4.1, this is not the case, and latitude and longitude coordinates are required in order for the spatial dimensionality of the observed and modelled data to match. If multiple stations were being evaluated, corresponding to a transect for example, the data was restructured to be in the form time by depth by distance. The distance of the station closest to shore was set to 0, and the relative distances from the previous station to the next was calculated using a haversine approximation.

3.2 GLORYS12v1

Model data is obtained from the GLORYS12v1 oceanographic reanalysis product (Lellouche, Greiner, Bourdalle-Badie, et al. 2021). GLORYS12v1 is a global ocean reanalysis product whose data are available on daily timescales on a 1/12° regular user grid. It is based on the NEMO3.1 ocean and sea ice model configuration with data beginning January 1, 1993. Its bathymetry is based on GEBCO8 for depths shallower than 200 m and ETOPO1 for depths deeper than 300 m, and a blending of the two in between. Surface forcings are obtained from the ERA-Interim atmospheric reanalysis until January 1, 2019 and ERA5 afterward. Runoff is prescribed to follow a climatological annual cycle from Dai et al. (2009), with prescribed open-ocean fluxes to account for glacier melt from Greenland and Antarctica (Lellouche, Greiner, Galloudec, et al. 2018). The majority of assimilated data comes from satellite products that constrain sea surface temperature, sea level anomaly, and sea ice concentration. In addition, in situ profiles are assimilated from the CORA4.1 database, which includes AZMP data. Thus, the comparisons performed here are not between independent data.

Temperature, salinity, and bottom temperature variables from GLORYS12v1 are analysed in the examples presented here. Prior to calling OMAtools in Example 1, a subset of GLORYS12v1 data surrounding the location of Stn27 was extracted, and interpolation was handled by OMAtools. In Example 2, the GLORYS12v1 fields were interpolated onto the AZMP Seal Island transect using a nearest-neighbour approach prior to calling OMAtools.

For Examples 3 and 4, the static model data for the ocean model bathymetry, the deepest valid temperature from the three-dimensional temperature field, and the GLORYS12v1 bottom temperature variable are extracted. For Example 4, bottom temperature is also presented on the original Arakawa C-grid that the model was run on.

4 Examples and Applications

Multiple facets of model skill are evaluated through the set of OMAtools metrics. The mean bias provides insight into how closely a model reproduces average observed conditions. This insight is particularly important in regards to state-dependent, non-linear processes like sea ice formation, whereby a model’s mean state can affect oceanographic tendencies in ways that can be amplified through internal feedbacks. Temporal correlations indicate how consistently models predict the tendencies of oceanographic variations relative to observations. Low correlations are likely an indication of missing or misrepresented processes or forcings. In contrast, normalized variance differences evaluate the magnitude of oceanographic variations, which can reflect the sensitivity of the system to forcings. Finally, gamma-squared combines aspects of both the correlation and the normalized variance differences to summarize whether the model represents oceanographic variations effectively enough to be considered a “useful” tool for the variables being examined. We interpret the results of the metric calculations to indicate model skill when i) biases are insignificant, ii) correlations are significantly above 0, iii) normalized variance differences are insignificant, and iv) gamma-squared is significantly lower than 1.

4.1 Example 1: comparing Stn27 temperature and salinity profiles

In this example, one-dimensional, time-varying comparisons are performed. Temperature and salinity profiles observed at AZMP Stn27 are provided as the reference dataset, and a subset of the gridded data from GLORYS12v1 data in the vicinity of the station forms the comparison dataset. Following TEOS-10 protocols (McDougall, Barker, et al. 2021), the GLORYS12v1 temperature and salinity outputs are interpreted to be conservative temperature and preformed salinity (scaled by a factor of $\text{ups}=(35.16504/35) \text{ g kg}^{-1}$), and the AZMP potential temperature and practical salinity are converted to conservative temperature and preformed salinity via the GSW Oceanographic Toolbox.

Prior to calling OMAtools, three modifications are made to the AZMP Stn27 profile data. First, the AZMP Stn27 profiles are vertically interpolated from the default CASTS pressure level coordinate (described in Section 3.1) to the 50 GLORYS12v1 depth levels. The reason for performing this vertical interpolation is that the vertical grid in CASTS is much higher resolution than the GLORYS12v1 vertical grid at most depths, yet there are many pressure levels where there is no AZMP data. **If this vertical interpolation were not performed before calling the OMAtools functions, the GLORYS12v1 data would have been automatically remapped to the CASTS vertical grid by the OMAtools functions,** and we prefer to avoid interpolating from a lower resolution to a higher resolution coordinate. This example provides an illustrative situation to explore the implications of input data choices on interpolation in more depth. An alternative option would have been to switch

the order of input datasets, making the GLORYS12v1 dataset the reference dataset and the Stn27 dataset the comparison dataset. In this particular case, where the Stn27 data consists of a single profile with latitude and longitude dimensions of length 1, this would have led to OMAtools attempting to interpolate it to the three-dimensional subset (depth by latitude by longitude) of the GLORYS12v1 data. This does not lead to the goal of performing a comparison at Stn27. However, if both datasets were truly three-dimensional, and there was no preference regarding which horizontal grid was used for the output, then the two datasets could be switched. This would have little effect on the bias (reversing the sign) and correlation metrics, but it could have subtle and important effects on the others. If the normalization option for the variance difference calculation were to normalize by the variance of the reference dataset, then both variance difference and γ^2 metrics involve a division by the variance of the reference dataset. Switching which dataset is considered the reference dataset will alter the results in a way not a priori predictable. Therefore, we choose here to perform the vertical interpolation prior to calling OMAtools functions.

Second, the OMAtools climatological bias and correlation calculations both require the two datasets to be matched along the time axis. As a result, AZMP data was averaged into daily bins to match the temporal frequency of GLORYS12v1 data, and only those days with AZMP data were retained from GLORYS12v1. Third, the spatial dimensionality of the Stn27 AZMP data (one-dimensional along depth) does not correspond to the dimensionality of the GLORYS12 data (three-dimensional with depth by latitude by longitude), which would generate an error in OMAtools. Consequently, dimensions of size one were added for latitude and longitude to the AZMP Stn27 data with coordinate values corresponding to the station location. This horizontal grid for the Stn27 data and the subset of grid points from GLORYS12v1 are not colocated, so this triggers spatial interpolation within OMAtools in this example.

The resulting climatological Stn27 temperature and salinity profiles, separated by season, are presented for GLORYS12v1 and AZMP in Figure 2. Plots of the climatological bias, correlation, normalized variance difference and gamma-squared are presented for conservative temperature and preformed salinity in Figure 3. The assessment of significance is performed for a significance value of 0.05 using a Bonferroni correction (see Section 2.3). Time-invariant summary metrics for the climatological bias, correlation, normalized variance difference and gamma-squared are presented in Table 1.

Metrics summarized over all depths in Table 1 indicate that temperature variations are generally better captured than salinity variations. Temperature correlations range between 0.58 to 0.98 over all seasons and depths, while salinity correlations range between -0.04 to 0.84. Also, mean gamma-squared values are lower for temperature than for salinity, and normalized variance differences are similar for both variables except during winter, when salinity variances are over-estimated much more substantially. Seasonally, GLORYS12v1 shows greater skill across all depths modelling salinity during spring and summer than winter, based on field-mean and field-RMSE values. In contrast, GLORYS12v1 shows less skill modelling temperatures during the summer than during the other seasons.

Examining metric results as a function of depth in Figure 3, GLORYS12v1 shows the highest skill in the top 25 m. Biases are generally insignificant, correlations are mostly higher than at other depths, variance differences are mostly insignificant and gamma-squared values are significantly lower than 1 during most seasons. These results are consistent with the

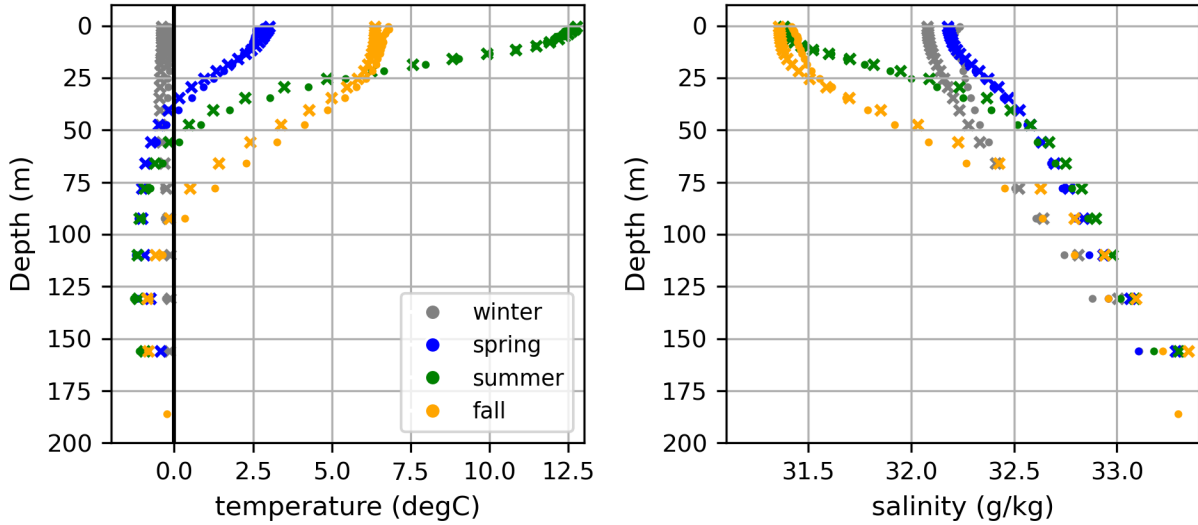


Figure 2: Time-averaged Stn27 temperature (left) and salinity (right) profiles for AZMP in dots and GLORYS12v1 in x's. A vertical black line in the temperature plot indicates 0°C. Seasons are indicated by dot or x colour, where winter (grey) corresponds to January-February-March, spring (blue) corresponds to April-May-June, summer (green) corresponds to July-August-September, and fall (orange) corresponds to October-November-December.

fact that a greater quantity of assimilated data is available at the surface than at other depths. During seasons when the sub-surface is well-mixed with the surface, this improved performance extends below the surface (e.g. to 50 m depth for winter temperatures). The one notable exception to model performance being better at the surface occurs for winter salinity, for which the most significant biases and highest gamma-squared values occur in the top 25 m of the water column. The source of such a discrepancy is not evident from the analyses performed here, but it could arise from the representation of sea ice formation and associated brine rejection upstream.

In contrast, GLORYS12v1 shows the least skill at predicting either temperature or salinity variations below 100 m depth at Stn27. Both temperature and salinity exhibit significantly positive biases there during all seasons except fall temperatures, which are biased significantly negatively between 100 m and 150 m depth. In addition to the biases, correlations are lowest below 100 m depth (particularly for salinity), and the variance is over-estimated there during most seasons. This corresponds to a gamma-squared that is near 1 for temperature or significantly higher than 1 for salinity. The seasonal climatologies at these depths indicate that oceanographic conditions change relatively little throughout the year, with temperatures below 0°C and salinities near 33 g kg⁻¹. These conditions are consistent with those of the cold intermediate layer (CIL) on the NL Shelf (Coyne, Cyr, Snook, et al. 2025). GLORYS12v1 has been previously shown to exhibit warm, salty biases near the bottom edge of the CIL and at the shelf-break on AZMP transects upstream of Stn27, which have been linked with on-shelf intrusions and misplacements of the shelf-break front (Andres, Soontiens, et al. 2024). It may be that conditions at Stn27 are being affected by these poorly

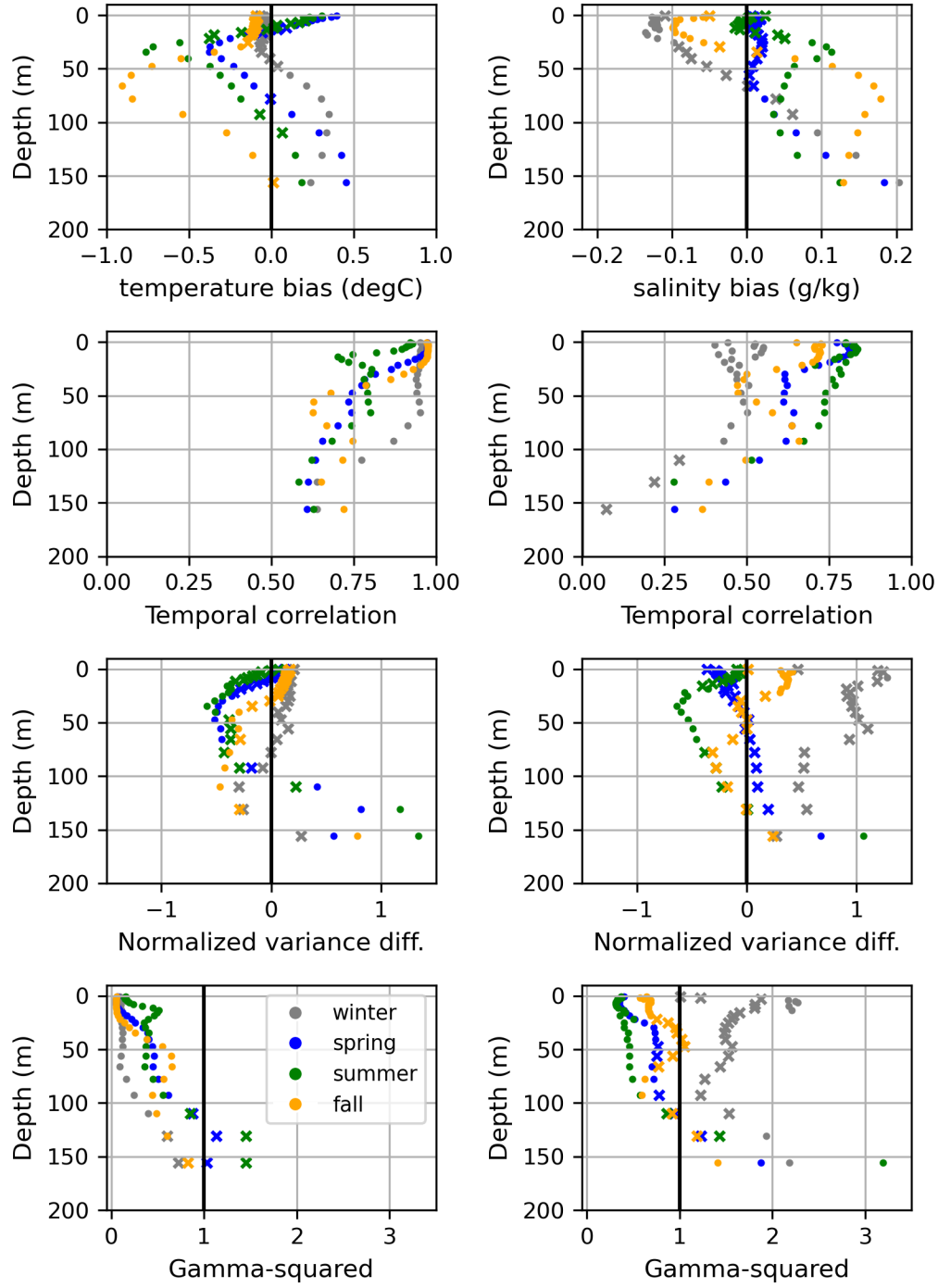


Figure 3: Metrics for temperature (left) and salinity (right), where filled dots are statistically significantly different than the reference, and x's are statistically insignificant values. The metrics include climatological bias (GLORYS12v1-AZMP, top row) as a function of depth at Stn27, with vertical reference lines at values of 0 (units dependent on subplot); temporal correlation (second row) with a reference value of 0; normalized variance difference (GLORYS12v1-AZMP, third row) with vertical reference lines at values of 0; and gamma-squared (bottom row) with vertical reference lines at values of 1. Colours indicate season as shown in the legend.

Variable	Metric	Season	fieldmin	fieldmax	fieldmean	field RMSE
Salinity	climBias (g/kg)	Winter	-0.14	0.20	-0.06	0.10
		Spring	0.00	0.18	0.03	0.03
		Summer	-0.02	0.12	0.04	0.05
		Fall	-0.10	0.18	0.00	0.10
	correlation	Winter	0.07	0.55	0.44	—
		Spring	0.28	0.81	0.69	—
		Summer	-0.04	0.84	0.72	—
		Fall	0.36	0.73	0.61	—
	varNormDiff	Winter	0.27	2.44	1.16	1.20
		Spring	-0.36	0.68	-0.08	0.18
		Summer	-0.64	1.06	-0.25	0.34
		Fall	-0.31	0.41	0.14	0.25
	gammaSqr	Winter	1.01	2.27	1.69	—
		Spring	0.35	1.88	0.63	—
		Summer	0.31	3.20	0.57	—
		Fall	0.58	1.41	0.80	—
Temperature	climBias (°C)	Winter	-0.07	0.35	0.05	0.11
		Spring	-0.38	0.46	0.05	0.25
		Summer	-0.76	0.31	-0.11	0.28
		Fall	-0.91	0.01	-0.27	0.28
	correlation	Winter	0.64	0.96	0.91	—
		Spring	0.61	0.98	0.84	—
		Summer	0.58	0.93	0.79	—
		Fall	0.63	0.98	0.86	—
	varNormDiff	Winter	-0.30	0.27	0.09	0.15
		Spring	-0.51	0.82	-0.11	0.29
		Summer	-0.59	1.34	-0.16	0.40
		Fall	-0.47	0.79	-0.02	0.22
	gammaSqr	Winter	0.08	0.73	0.17	—
		Spring	0.05	1.14	0.32	—
		Summer	0.14	1.45	0.45	—
		Fall	0.05	0.83	0.26	—

Table 1: Time-invariant metrics related to comparisons at Stn27 in Example 1. CASTS includes 90 profiles within the years 1993-2024 during the winter, 392 profiles during the spring, 216 profiles during the summer, and 347 profiles for temperature and 348 for salinity during the fall.

constrained upstream processes.

At intermediate depths (approximately 25 m to 100 m), GLORYS12v1 shows skill in representing temporal variations for all seasons, but that performance is better for temperature than for salinity. There are significant cold biases at depths between 25 m to 75 m from spring until fall, with the most biased conditions found in the fall. Combined with the positive temperature biases in the top 10 m during spring and summer, this may be an in-

dication of an under-estimation of vertical mixing. Yet, salinity biases at those depths are mostly insignificant. Alternatively, these biases may arise from the different temporal characteristics of the instantaneous observations, which can be influenced by diurnal variations, versus daily-averaged model data. Temperature shows ranges of depths with degraded performance, where correlations are lower and gamma-squared values are higher than values just above and below. The depths where these instances of reduced performance occur vary by season: 25 m to 50 m during spring, approximately 15 m for summer, and 50 m to 70 m during fall. The temperature variance is also under-estimated at these depths, although not over as narrow a depth range. These depth ranges correspond to strong gradients in the climatological temperatures in Figure 2, so small variations in the depths or properties of water masses could lead to large differences between the model and observations. It is notable that there are no corresponding decreases in performance for the winter, which does not exhibit similarly strong vertical gradients. Such reductions in performance are apparent for salinity during the spring, summer, and fall, although the magnitude and vertical extent is much smaller in the summer.

4.2 Example 2: comparing measurements and model on the Seal Island transect

For the second case, data on the Seal Island (SI) transect is compared between AZMP data and GLORYS12v1. The preparation of AZMP profiles from the CASTS dataset into a two-dimensional transect dataset with dimensions of time, depth and distance is described in Section 3.1. GLORYS12v1 data is remapped using a nearest neighbour approach to the SI transect stations prior to calling OMAtools, as the dimensions of the two input datasets do not match otherwise (depth by distance versus latitude by longitude). Such a mismatch would trigger an error in OMAtools.

The SI transect is only occupied as part of AZMP twice a year, during summer and fall, and there are 19 and 12 occupations in these seasons, respectively, between 1993 and 2024. In addition, partial occupations at least one station of the SI transect are included in the comparisons here, making up a total of 31 observations during the fall and 69 temperature (67 salinity) observations during summer. This is a much smaller number of data points upon which to make comparisons than for Stn27.

Temperature and salinity climatologies separated by season are presented in Figure 4. Key features of SI summer conditions are present in both GLORYS12v1 and the observations. These are the presence of the CIL underlying a warm surface layer and strong lateral temperature and salinity gradients at the shelf break. GLORYS12v1 appears to under-estimate the strength of both shelf-break gradients. The CIL has effectively disappeared by fall, and the surface fresh layer observed during the summer has been mixed into deeper layers for most of the shelf.

Examining the climatological bias plots for summer in the top row of Figure 5, the weaker shelf-break gradients appear as significantly positive temperature and salinity biases just inshore of the shelf-break. These conditions suggest that GLORYS12v1 may permit too much mixing across the shelf-break front, as noted in Andres, Soontiens, et al. (2024). There are also positive temperature biases around 20 m depth across most of the shelf, which are co-

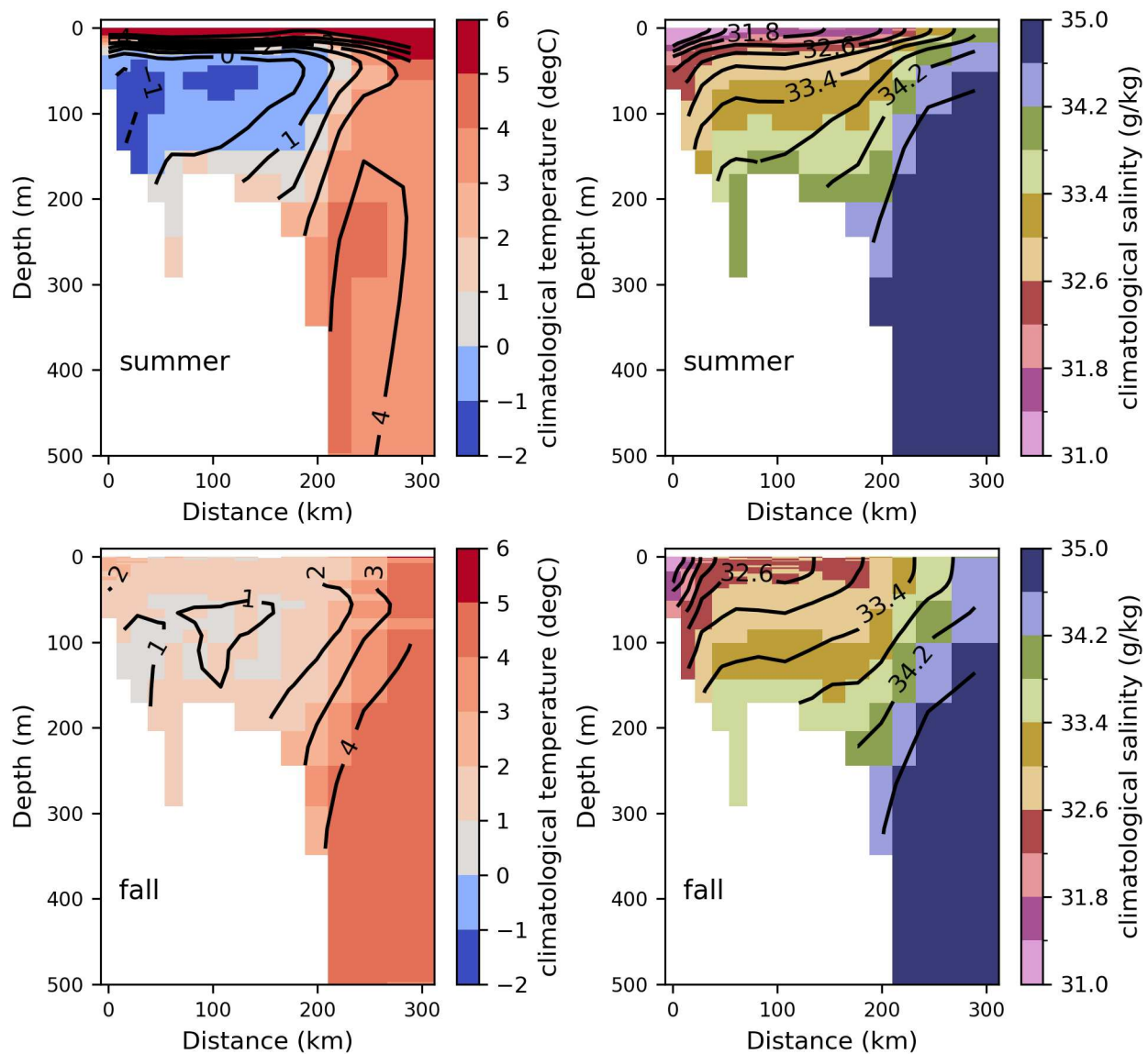


Figure 4: Temperature (left) and salinity (right) climatologies on the SI transect for summer (top row) and fall (bottom row). CASTS data is plotted in colours, and GLORYS12v1 data is plotted in black contours for the same levels. Dashed contours indicate negative values.

Variable	Metric	Season	fieldmin	fieldmax	fieldmean	field RMSE
Salinity	climBias (g/kg)	Summer	-0.41	1.23	0.18	0.33
		Fall	-0.60	0.58	0.07	0.27
	corr	Summer	-0.10	0.86	0.44	—
		Fall	-0.93	1.00	0.52	—
	varNormDiff	Summer	-0.80	3.68	0.37	1.01
		Fall	-0.85	27.63	1.78	3.34
	gammaSqr	Summer	0.26	4.66	1.36	—
		Fall	0.10	18.40	1.95	—
Temperature	climBias (°C)	Summer	-0.86	2.56	0.51	0.78
		Fall	-1.00	0.87	0.13	0.33
	corr	Summer	-0.16	0.98	0.60	—
		Fall	0.01	1.00	0.82	—
	varNormDiff	Summer	-0.85	3.07	-0.20	0.57
		Fall	-0.81	9.84	0.44	1.34
	gammaSqr	Summer	0.15	2.58	0.74	—
		Fall	0.00	2.75	0.43	—

Table 2: Time-invariant metrics related to comparisons at Seal Island in Example 2. There are 31 transect observations of temperature and salinity during fall and 69 temperature transects and 67 salinity transects during summer.

located with strong vertical gradients in the summer temperature plot in Figure 4. Salinity biases at this depth are insignificant and negative except for a patch just off the shelf. These two features may suggest an over-estimation of vertical mixing across the thermocline and halocline, but the absence of a corresponding cold anomaly near the surface in GLORYS12v1 warrants further investigation. Finally, there exist localized, significantly positive salinity biases at the bottom near the inshore slope. As a whole, the field-aggregated statistics in Table 2 indicate that mean summer biases for both temperature and salinity are positive.

The climatological temperature biases are insignificant in the fall in Figure 6, although the positive salinity biases remain near the shelf-break and along the inner slope.

Interannual correlations are mostly insignificant for temperature and particularly for salinity during the summer in Figures 5. GLORYS12v1 shows greater skill at predicting interannual summer variations off the shelf than on the shelf. On the continental shelf, correlations are higher in the inshore than elsewhere except for the top 5 m, where the correlations are significantly positive. These high correlations near the surface are likely due to the data assimilation, which includes satellite-based SST data products. In Figure 6, interannual temperature correlations during the fall are significantly positive in the top 20 m across the entire shelf. The reduced stratification and increased vertical mixing in the fall makes surface variations more representative of sub-surface conditions during the fall than the summer. The increase in temporal correlations for temperature from summer to fall is notable in Table 2, where the field minimum, field maximum and field mean all increase. As in the previous example, salinity correlations are generally lower than temperature.

Finally, there is insufficient data to extract many significant metric values for normalized variance difference and gamma-squared during either season in Figures 5 and 6. Yet, the

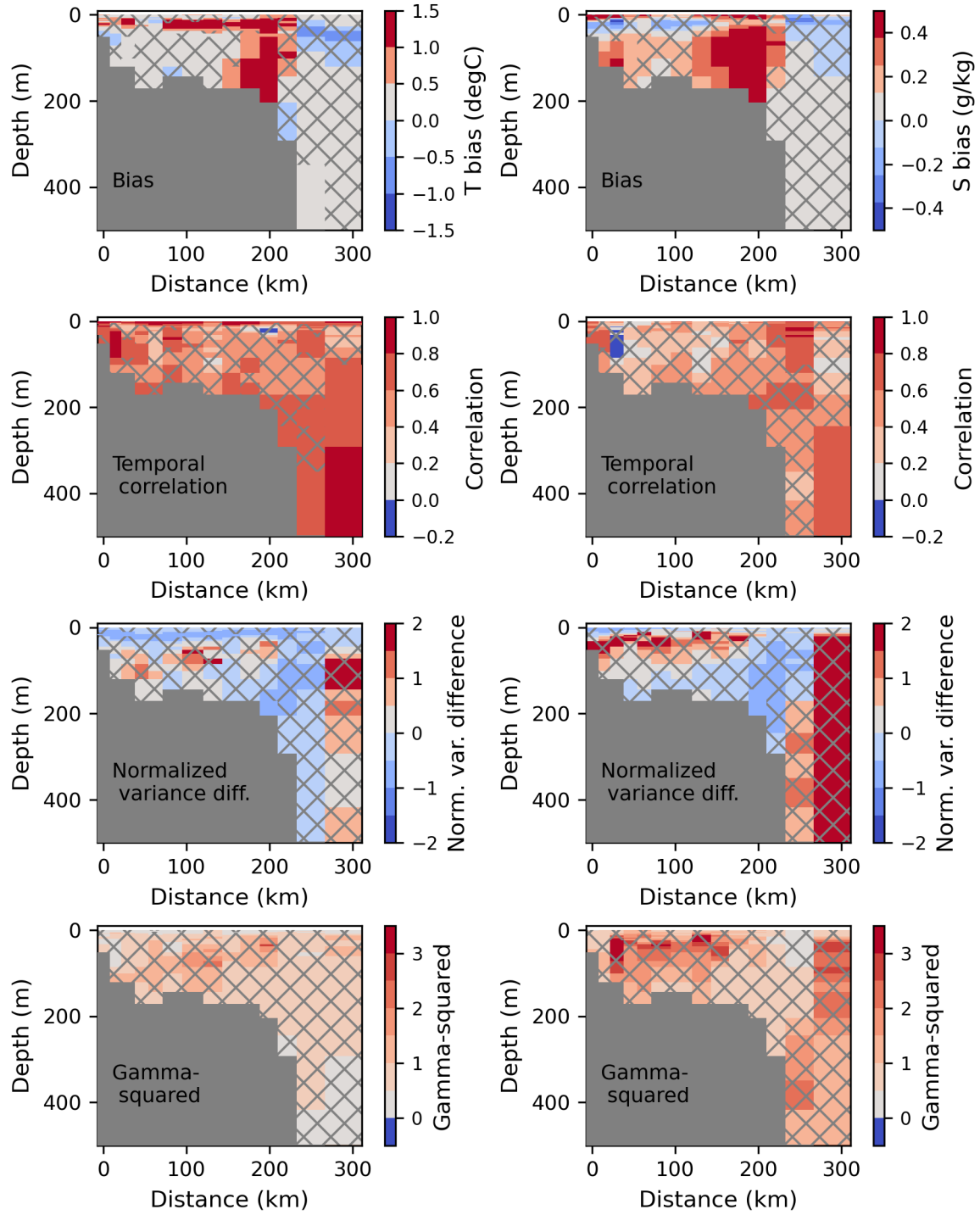


Figure 5: Temperature (left) and salinity (right) metrics for the SI transect for the summer (July-August-September). The metric is labelled on each plot, and grey hatching indicates values insignificantly different than 0 for climatological bias (units either $^{\circ}\text{C}$ or g/kg), temporal correlation and normalized variance difference (both unitless). Insignificant values for gamma-squared are relative to a value of 1.

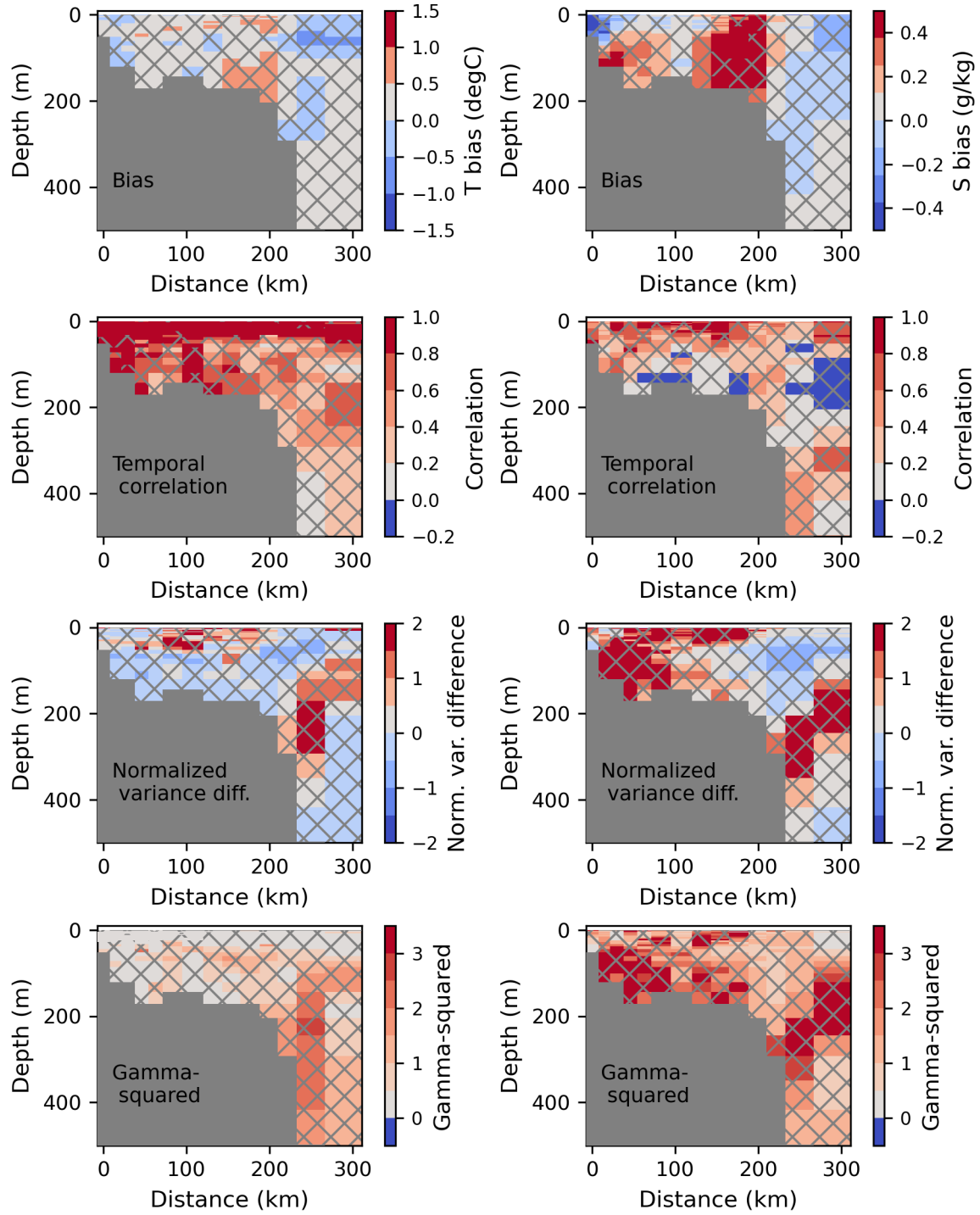


Figure 6: Temperature (left) and salinity (right) metrics for the SI transect for fall (October-November-December). The metric is labelled on each plot, and grey hatching indicates values insignificantly different than 0 for climatological bias (units either $^{\circ}\text{C}$ or g/kg), temporal correlation and normalized variance difference (both unitless). Insignificant values for gamma-squared are relative to a value of 1.

model shows significant skill in predicting temperature variations around 20 m depth during the fall, particularly in the inshore. When aggregated over the field, GLORYS12v1 shows greater model skill during the summer than fall based on these two metrics for salinity, and approximately equivalent performance for the two seasons for temperature.

4.3 Example 3: comparing GLORYS12v1 user-grid bathymetry against the depth of the deepest temperature

For the third and fourth examples, different representations of model bathymetry and bottom temperature on the NL Shelf are compared. The motivation behind this comparison is that differences have been noted between GLORYS12v1 bottom temperature fields during previous analyses, and the authors have been unable to locate documentation to indicate how “bottomT” was calculated. The model simulations underlying the GLORYS12v1 data product are not generated on the regular user grid, but on a curvilinear Arakawa C-grid configuration in the horizontal with a partial step representation for the ocean bottom. “bottomT” exists as a two-dimensional variable on the C-grid as well as the user grid. If the user-grid “bottomT” were horizontally interpolated from the two-dimensional “bottomT” C-grid field, it may have errors where the bathymetry varies, due to conditions around a given point being treated as equivalent in influencing the point, when some may actually be located at different depths. On the other hand, remapping the three-dimensional temperature field from the C-grid to the user grid can also lead to differences at the bottom due to differences in the representation of the bathymetry on the two grids, which may affect “deepestT”. Hence, in this example, we explore bathymetry differences on the user grid and in Example 4, we explore the bottom temperature differences themselves.

Time-independent comparisons do not provide statistical assessments of the significance of the metrics provided. As a result, more caution needs to be made when interpreting their results. Here, the time-independent OMAtools functions are applied to the different representations of bathymetry on the user grid: that released in the model statics file and the depth of the bottom edge of the deepest grid cell containing valid data. These time-invariant variables are ideally suited for time-independent comparisons. Figure 7 shows the static fields of bathymetry and depth to the bottom of the deepest level on the user grid.

fieldmin (m)	fieldmax (m)	fieldmean (m)	field RMSE (m)
-350	190	-30	61

Table 3: Time-invariant metrics related to bathymetry differences in Example 3.

Differences are mostly small between the two bottom depth plots. A spatial regression of the depth of the bottom level against the bathymetry field has an R-squared value of 1.00 and a slope of 0.98866 ± 0.00006 , which indicates that the spatial patterns are nearly identical. The plot of the differences themselves in Figure 7 shows alternating patterns of large and small, negative depth differences in the central Labrador Sea and Labrador slope. When aggregated as a whole over the entire field, the mean difference is -30 m. Thus, the bathymetry field is deeper than the depth to the bottom of the deepest level with valid data. In contrast, differences are much smaller in magnitude on the continental shelf. This result

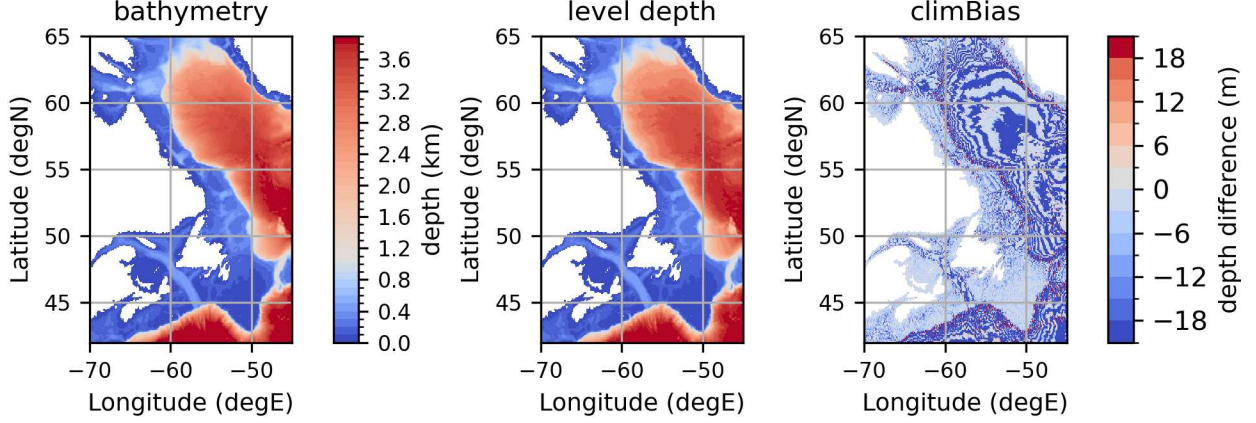


Figure 7: Static bathymetry field on the user grid (left), the depth of the deepest user grid model level with valid data (middle), and the difference between the two (right).

is as expected, since the deepest levels are orders of magnitude thicker than the shallower levels.

4.4 Example 4: comparing different representations of modelled bottom temperatures on the NL Shelf

For the final example, two different GLORYS12v1 variables are compared: the “bottomT” variable, which is released as a two-dimensional variable with GLORYS12v1 user-grid output, and the deepest valid temperature value (“deepestT”) extracted from the three-dimensional GLORYS12v1 temperature field on the user grid. Since both datasets are from GLORYS12v1, the comparisons are based on complete daily time series from 1993 to 2024 with complete spatial coverage. This results in a high number of degrees of freedom. Also, since there is no observational data in this case, “bottomT” is arbitrarily chosen to be the reference dataset and “deepestT” to be the comparison dataset.

Climatologies for “bottomT” on the C-grid, “bottomT” on the user grid and “deepestT” on the user grid are presented in Figure 8 based on years 1993 to 2019 (the latest year of C-grid data available at the time of calculation). Differences are small between these three variables, but they are evident in the central Labrador Sea (at approximately 57°N, -55°E), where both “bottomT” variables appear similarly colder than “deepestT”.

Metric	fieldmin	fieldmax	fieldmean	field RMSE
climBias (°C)	-5.40	3.19	0.01	0.10
corr	-0.37	1.00	0.96	—
varNormDiff	-0.98	36.89	0.02	0.21
gammaSqr	0.00	35.15	0.09	—

Table 4: Time-invariant metrics related to bottomT and deepestT comparisons in Example 4.

Comparison metrics for “bottomT” and “deepestT” on the GLORYS12v1 user grid are

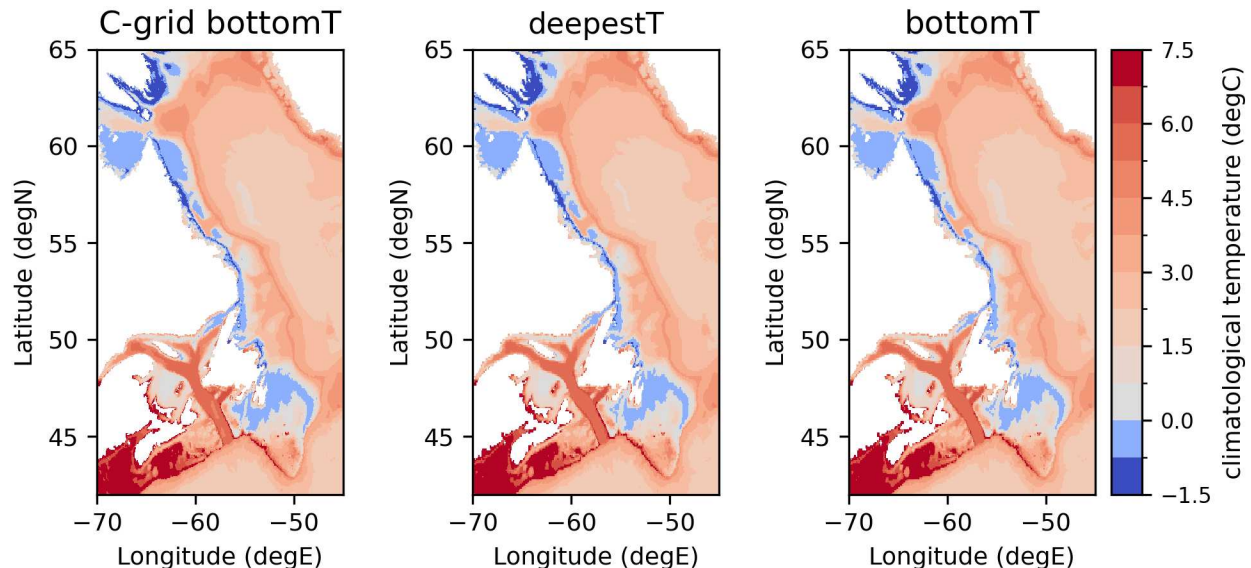


Figure 8: Climatologies of bottom temperature on the C-grid (left), deepest temperature based on the user-grid three-dimensional temperature field (centre), and bottom temperature on the user grid (right) from GLORYS12v1 for all days from 1993 to the end of 2019.

presented in Figure 9 and Table 4. Climatological biases are significant in most grid cells, even though they are not always strong. This likely indicates that the biases are sustained over much of the time series. “deepestT” is warmer than “bottomT” in the deepest parts of the Labrador Sea with alternating regions of strong and weak biases aligned with bathymetric differences in this area (see Section 4.3 and Figure 7). On the continental Newfoundland and Labrador Shelf, biases tend to be of the opposite sign and strongest in locations of strong bathymetric gradients. When aggregated over the entire field, climatological biases are near 0°C.

Correlations are significantly positive over the entire domain, with a mean value over the entire field of 0.96 in Table 4 and values uniformly close to 1.0 on the continental shelves. In the deepest slopes and central Labrador Sea, correlations are the weakest. These regions are also the only locations where normalized variance differences are significant, with “deepestT” varying less than “bottomT” on the slopes and more than “bottomT” in the central Labrador Sea. On the whole, the combination of these results leads to gamma-squared values that are significantly less than 1 over most of the domain, but particularly so over the continental shelves. We conclude that “bottomT” and “deepestT” vary similarly on the continental shelves. In contrast, the representations are less consistent on the deep slopes and in the central Labrador Sea.

5 Conclusions

OMAtools is a python package with functions that can provide quick comparisons of two oceanographic datasets including assessments of significance. The metrics cover the range of key data attributes while remaining general as to variable or geometry. **Built-in interpola-**

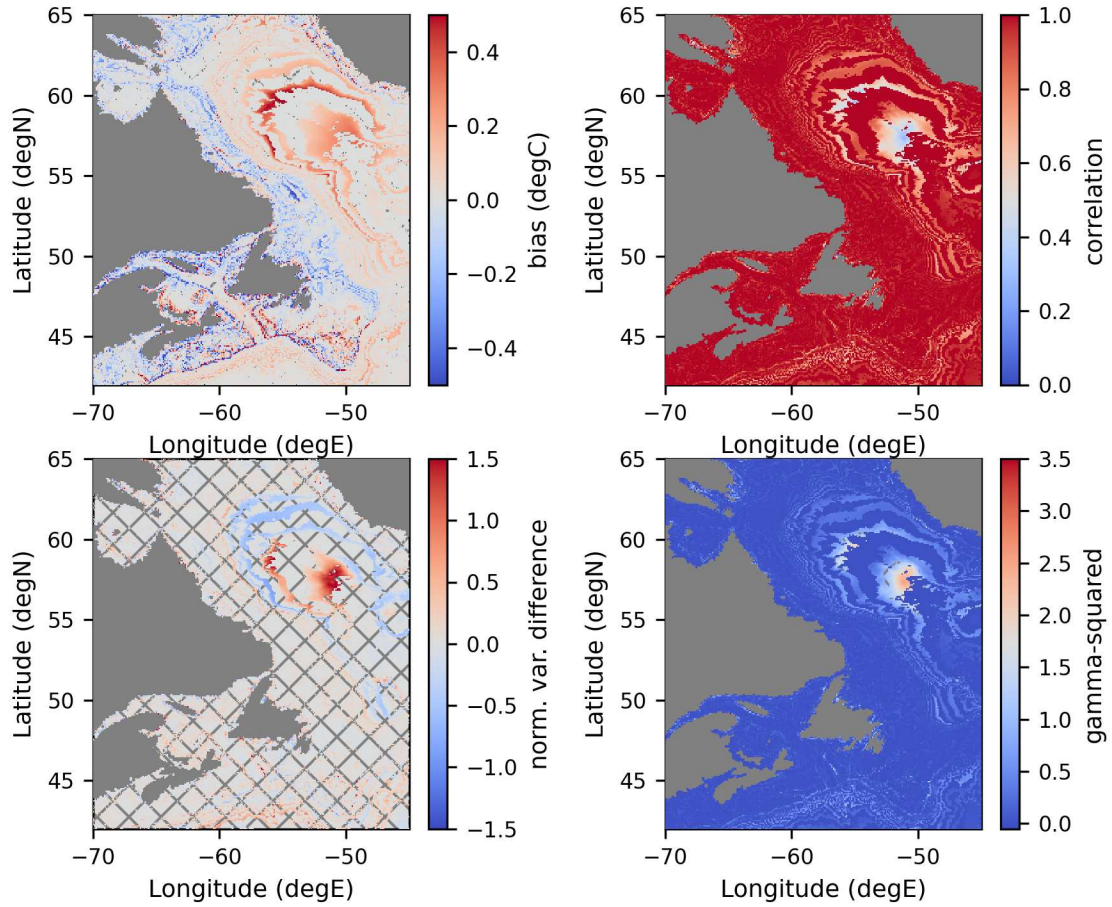


Figure 9: (Top left) Climatological bias for “deepestT” minus “bottomT”, (top right) correlation with time, (bottom left) normalized variance difference, and (bottom right) gamma-squared for “bottomT” and “deepestT” on the user grid.

tion tools permit users to perform as little preparation of the data as they would like, except when using curvilinear or irregular grids, which are not handled by the interpolation scheme. These limitations and others are flagged in the methods descriptions. It is recommended that users consider the implications of data preparation choices carefully.

This report presents how the data comparison metrics and their significance are calculated, as well as the assumptions inherent in these calculations. Assumptions are also described in the docstrings of the tools themselves. One highlighted point is that all significance assessments have been performed independently in each grid cell, which may not be appropriate if there are cross-correlations in the data. It is left to the user to interpret the meaning of results and the appropriateness of their applications.

Four examples of applications of the OMAtools functions are presented here. The first compares one-dimensional profiles at Stn27 from AZMP data and data surrounding that point from the GLORYS12v1 model. This example shows that the model reproduces temperature variations better than salinity, but shows skill in both for depths shallower than 100m. Biases are evident for both variables that are not uniform with depth. Example 2 presents comparisons between GLORYS12v1 and AZMP data on the Seal Island transect, which is a two-dimensional field of depth by cross-shelf distance. These comparisons are performed separately for summer and fall. Both temperature and salinity show significant positive biases in the outer shelf near the shelf-break during the summer. Similar but mostly insignificant biases are evident during the fall. The model shows significant skill (based on gamma-squared) in predicting fall temperature variations in the top 20 m in the inshore. Example 3 presents a time-invariant comparison of GLORYS12v1 bathymetry versus the depth of the deepest grid cells with valid data from the three-dimensional temperature field on the user grid. Alternating patterns of strong and weak biases are evident in the deepest regions of the Labrador Sea, with a net effect that the depth of the deepest cell under-estimates the bathymetry. Example 4 presents an evaluation of two different representations of bottom temperature in the vicinity of Newfoundland and Labrador from GLORYS12v1. These fields are two-dimensional (latitude by longitude). These variables provide nearly equivalent representations of bottom temperature variations over the continental shelf, but show significant biases along continental slopes and particularly in the deepest regions of the Labrador Sea where bathymetric differences were shown to be largest in Example 3. These examples illustrate the range of capabilities available within OMAtools that can be used to kick off more in-depth analyses or for a quick evaluation.

Near-future anticipated developments of OMAtools include adding more informative metadata to function outputs and resolving the limitation requiring an interpolation in time when interpolating in space for time-varying datasets. In the longer term, additional capabilities plan to be included in OMAtools. Two such capabilities will include trend analyses on biases and metrics for vector fields. The former will provide insight into whether biases are growing or shrinking, which can be helpful when assessing their cause. The latter will expand the types of variables for which OMAtools assessments can be applied.

6 Code Availability

OMAtools is accessible on gitlab at <https://gitlab.com/dfo-nl-oceanography/omtools.git>. Function descriptions are consistent with v1.2.0, and all calculations were performed with this release.

7 Acknowledgments

The authors thank Stephanie Taylor and Rick Danielson for their constructive reviews.

References

- Andres, H. J., Soontiens, N., et al. 2024. Seasonal variations of the cold intermediate layer on the Newfoundland and Labrador Shelf. *Progress in Oceanography* **229**: 103379. ISSN: 0079-6611. 10.1016/j.pocean.2024.103379.
- Castillo-Trujillo, A. C. et al. 2023. An evaluation of eight global ocean reanalyses for the Northeast U. S. Continental shelf. *Progress in Oceanography* **219**: 103126. ISSN: 0079-6611. 10.1016/j.pocean.2023.103126.
- Coyne, J., Cyr, F., Atchison, S., et al. 2026. More than a century of oceanic hydrography observations reveals profound climate-related changes in the Northwest Atlantic Ocean and its Arctic gateways. *Earth System Science Data* **18**: 1463–1488. 10.5194/essd-18-1453-2026.
- Coyne, J., Cyr, F., Donnet, S., et al. 2025. *Canadian Atlantic Shelf Temperature-Salinity (CASTS)*. date accessed: July 22 and 23, 2025. 10.20383/103.01191.
- Coyne, J., Cyr, F., Snook, S., et al. 2025. *Physical oceanographic conditions on the Newfoundland and Labrador Shelf during 2024*. Canadian Technical Report of Hydrography and Ocean Sciences: iv + 61p. 10.60825/1d14-x857.
- Dai, A. et al. 2009. Changes in continental freshwater discharge from 1948 to 2004. *Journal of Climate* **22**: 2773–2792. 10.1175/2008JCLI2592.1.
- Filliben, J. J., Heckert, A. 2012. Levene Test for equality of variances. In: *NIST/SEMATECH e-handbook of statistical methods*. Ed. by C. Croarkin, P. Tobias. Accessed: 2025-11-24. National Institute of Standards and Technology. 10.18434/M32189.
- Griffies, S. M. et al. 2009. Coordinated Ocean-ice Reference Experiments (COREs). *Ocean Modelling* **26**(1): 1–46. ISSN: 1463-5003. <https://doi.org/10.1016/j.ocemod.2008.08.007>.
- Johnson, R. A., Wichern, D. W. 2007. *Applied Multivariate Statistical Analysis*, 6th Ed. In: Pearson Education, Inc. Chap. Inferences about a mean vector.
- Lellouche, J.-M., Greiner, E., Bourdalle-Badie, R., et al. 2021. The Copernicus global 112° oceanic and sea ice GLORYS12 reanalysis. *Frontiers in Earth Science* **9**: 698876. 10.3389/feart.2021.698876.
- Lellouche, J.-M., Greiner, E., Galloudec, O. L., et al. 2018. Recent updates to the Copernicus Marine Service global ocean monitoring and forecasting real-time 1/12° high-resolution system. *Ocean Science* **14**: 1093–1126. 10.5194/os-14-1093-2018.
- McDougall, T. J., Barker, P. M. 2011. *Getting started with TEOS-10 and the Gibbs Seawater (GSW) Oceanographic Toolbox*. May. SCOR/IAPSO WG127: p. 28. ISBN: 978-0-646-55621-5.

- McDougall, T. J., Barker, P. M., et al. 2021. The interpretation of temperature and salinity variables in numerical ocean model output and the calculation of heat fluxes and heat content. *Geoscientific Model Development* **14** (10): 6445–6466. 10.5194/gmd-14-6445-2021.
- NumPy Developers. 2008-2025. *numpy.interp*. <https://numpy.org/doc/stable/reference/generated/numpy.interp.html>. Accessed: 2026-04-15.
- Ryan, A. et al. 2015. GODAE OceanView Class 4 forecast verification framework: global ocean inter-comparison. *Journal of Operational Oceanography* **8** (sup1): s98–s111. ISSN: 1755-876X. 10.1080/1755876X.2015.1022330.
- The SciPy community. 2008a. *scipy.interpolate.griddata*. <https://docs.scipy.org/doc/scipy/reference/generated/scipy.interpolate.griddata.html>. Accessed: 2026-04-15.
- 2008b. *scipy.interpolate.RegularGridInterpolator*. <https://docs.scipy.org/doc/scipy/reference/generated/scipy.interpolate.RegularGridInterpolator.html>. Accessed: 2025-11-24.
- Therriault, J.-C. et al. 1998. *Proposal for a Northwest Atlantic Zonal Monitoring Program*. Canadian Technical Report of Hydrography and Ocean Sciences: vii + 57 p.
- Thompson, K. R. et al. 2003. Prediction of surface currents and drifter trajectories on the inner Scotian Shelf. *Journal of Geophysical Research* **108**: 3287. 10.1029/2001JC001119.
- Weitzel, N. et al. 2024. Towards spatio-temporal comparison of simulated and reconstructed sea surface temperatures for the last deglaciation. *Climate of the Past* **20**: 865–890. 10.5194/cp-20-865-2024.