Development of Mechanistic Models

Assessment of model performance

Miljø- og Fødevareministeriet
Miljøstyrelsen

Technical Note
February 2019

The expert in **WATER ENVIRONMENTS**

Development of Mechanistic Models

Assessment of Model Performance

Prepared for	Danish EPA (Miljøstyrelsen, Fyn)
Represented by	Mr. Harley Bundgaard Madsen, Head of Section

Eelgrass in Kertinge Nor

Photo: Peter Bondo Christensen

Project manager	Anders Chr. Erichsen & Mads Birkeland
Quality supervisor	Anne Lise Middelboe

Project number	11822245
Approval date	19-02-2019
Revision	Final version : 1.0
Classification	Open

DHI • Agern Allé 5 • • DK-2970 Hørsholm • Denmark
Telephone: +45 4516 9200 • Telefax: +45 4516 9292 • dhi@dhigroup.com • www.dhigroup.com

CONTENTS

TABLES

1 Introduction

As part of the preparation towards the Danish River Basin Management Plans 2021-2027 the Danish Environmental Protection Agency (EPA) has initiated a number of mechanistic model developments with the aim of increasing the spatial coverage of models, improving the calibration/validation and hence the confidence of the Maximum Allowable nutrient Inputs (MAIs).

The development of mechanistic models includes both development of hydrodynamic models (water levels, currents, salinity, water temperatures, etc.) and biogeochemical models (nutrients, phytoplankton, organic matter, benthic vegetation, etc.). The model development entails several different processes (and sub-processes):

- Collection and preparation of input data (boundaries, meteorological forcings, initial data, etc.)
- Collection and preparation of observational data (monitoring) for model calibration/validation
- The calibration (adjustment of model parameters to minimize deviations between observations and model predictions)
- The validation (comparison with independent observations)

For all models (hydrodynamic and biogeochemical) the different processes are equally important, but with respect to the future model use especially the calibration and validation are very important. Through the calibration and validation, the quality of the models is evaluated, and can thus provide an indication of the quality of model predictions. The latter will support scenario modelling and eventually support the estimations of MAI, which is the overarching aim of the entire model development.

This technical note describes the formalized validation procedure which we will apply throughout the development of the mechanistic models.

2 Calibration and Validation

As stated earlier a number of mechanistic models are being developed as part of the model development towards the Danish RBMP 2021-2027. Each mechanistic model consists of a hydrodynamic model seamlessly coupled to a biogeochemical model (for details of the different models we refer to DHI (2017a) and DHI (2017b)). The models being developed consist of two regional models, three local models and six estuary specific models:

- Regional models
    - A model covering the inner Danish waters (IDF-model)
    - A model covering the North Sea (NS-model)
- Local models
    - A model covering the northwestern Belt Sea
    - A model covering the Little Belt area
    - A model covering the Smålandsfarvand and adjacent waters

- Estuary specific models
    - Ringkøbing Fjord
    - Nissum Fjord
    - Limfjorden
    - Mariager Fjord
    - Odense Fjord
    - Roskilde Fjord

All in all, the model development project includes 11 separate model setups, model calibrations and model validations. Development of individual models will follow the same rigorous scheme, and every process in model development will be fully documented. Main activities in model development include:

- **Model setup:** Model setup includes the collection and preparation of input data (boundaries, meteorological forcings, initial data, etc.) and the collection and preparation of observational data for calibration/validation. When data have been collected and prepared according to the model structures (see DHI (2017a) and DHI (2017b) for details) the different data are combined in MIKE 3 FM and MIKE 3 ECO Lab, which is the modelling software applied for this specific modelling study.

- **Model calibration:** Following the model setup, the different models are calibrated. This entails the process where model parameters are adjusted to allow for the best fit between modelled parameters and similar observed parameters. Observations are used intensively, during this process, and encompass observed water levels, salinities, water temperatures and biogeochemical parameters, like observations of nutrients, chlorophyll-a, light attenuation coefficients (K_d) and oxygen concentration.

- **Model validation:** Finally, the calibrated models are validated. Validation, at best, covers a process where model results are compared to independent datasets. These independent datasets could be additional years, not used for the calibration, or supplementary monitoring stations. Not all models (or water bodies) have enough observational data to carry out a validation and for those water bodies the calibration and validation will be merged, but we aim at doing a prober validation in as many models and water bodies as possible.

According to the model development plans, we aim at running all models for the period 2002-2016 providing 15 years of model results. Due to the amount of years modelled we aim to use the period prior to 2011 as calibration period, and the years 2011-2016 as our validation period. When at a later stage the models will be applied for scenario modelling, the last 5-6 years will be used to assess effects of e.g. reducing nutrient loadings, and thus the scenario results will coincide with the validation period.

To allow for a robust and objective analysis of the model quality (validation) a small literature survey was carried out reviewing a number of different metrics (goodness-of-fit) for evaluating the model quality. This review resulted in three goodness-of-fit measures which we will apply during the model developments, and these metrics are condensed in the following section.

3 Goodness of Fit

If ecosystem models are used to guide the management of water bodies, models should be reliable in a broad sense and able to reproduce observational data. To this end, numerous metrics were developed and applied to numerically quantify the *goodness-of-fit* (or skill) of models to observational data. Different skill metrics assess different aspects of model performance and several metrics are needed to analyse the performance of an ecosystem model. However, among the large pool of available skill metrics (see Bennett *et al.* 2013, Moriasi *et al.* 2012, 2015, Stow *et al.* 2009) many metrics are redundant because they reflect the same aspects of skill performance, e.g. the group of "error indices" Mean Absolute Error (MAE), Mean Square Error (MSE) and Root Mean Square Error (RMSE) are all measures of *average error* between an observation and a model prediction (Olsen *et al.* 2016), but without indicating if the model over- or under-estimate observations.

A closely related metric is "Percentage Bias" (P-Bias) that expresses a normalized error value by dividing the summed difference between observations and model prediction with the sum of observational data. All error metrics score from 0 (best – perfect match) to indefinite (negative or positive). When assessing model skills, we will use P-bias as a metric reflecting to what extent the model represents a sufficiently correct level (e.g. concentration of nutrients, chlorophyll-a) compared to observations.

Another category of skill metrics is "correlation" indices including four variants of correlation: Spearman, Pearson, Kendall, and coefficient of determination (R^2). R^2 is oversensitive to high extreme values, but insensitive to level differences between model predictions and measured data (Legates & McCabe 1999, Krause *et al.* 2005). We will use the non-parametric Spearman rank correlation coefficient which is slightly forgiven if peak observations and model predictions (e.g. spring bloom) are mistimed by 1-2 weeks.

The most universal metric is the Modelling Efficiency Factor (MEF) that is closely related to the Nash-Sutcliffe Model Efficiency (Nash & Sutcliffe 1970), which is a measure of the ratio of the model error to the variability of the observational data. The metric was originally developed to assess the performance of river catchment models, which exhibit a similar seasonal variability to phytoplankton and inorganic nutrients (rapid increases and decreases).

Below we list the suite of model performance metrics which we will use, highlighting their advantages and limitations.

Skill metric	Equation	Remarks
Percent Bias	$P_{bias} = \frac{\sum_1^N (P_i - O_i)}{\sum_1^N O_i} * 100$	Percent model-bias (P_{bias}) is expressed by summed differences between modelled (P) and observed values (O) normalized to the sum of observations. P_{bias} shows if the model systematically over- or underestimates observations (O); the closer model predictions (P) are to "0" the better representation of data. P_{bias} increases the weighting of errors relating to low measurement values (e.g., low summer concentrations of phytoplankton and inorganic nutrients); hence, insufficient model-tracking of low value observations will invariable result in high P_{bias}. The amplification of errors during periods when low parameter values (observational and modelled) dominate can be overcome by splitting seasonally variable data into two sections, e.g. representing" growth season" (April through September) and "winter" (December through February). P_{bias} and the related error metrics cannot reflect trends or temporal variation in data.

Skill metric	Equation	Remarks
Spearman Rank correlation	$r_s = 1 - \frac{6\sum_1^N (rg_{Oi} - rg_{Pi})^2}{N(N^2-1)}$	Spearman rank correlation For a sample of size N, the N raw scores of observations O_i and model predictions P_i are pair-wise converted to ranks rg_{Oi} and rg_{Pi} and the differences subsequently calculated, summed, multiplied by 6 and divided by the denominator (see equation). Typically, calculation of Spearman rank correlation is carried out using statistical programs that also calculate r_s and the associated level of significance. The Spearman rank correlation is less sensitive than the Pearson correlation, R and R^2 to strong outliers that occur in the tails of both samples. That is because Spearman's rho limits the outlier to the value of its rank.
Modelling Efficiency Factor	$MEF = 1 - \frac{\sum_1^N (P_i - O_i)^2}{\sum_1^N (O_i - \bar{O})^2}$ $= 1 - \left(\frac{RMSE}{SD_O}\right)^2$	Modelling Efficiency Factor (MEF) is expressed by the squared sum of differences between modelled (P) and observed values (O) normalized to the standard deviation of observations. MEF is the only metric selected that assesses both model precision and accuracy. Among the metrics used MEF is both most sensitive to scale-off effects and inverse relations, but less sensitive to lack of correlation or mismatch of trends. In that sense, MEF and Spearman Rank correlation metrics supplement each other. An error-free model attains a MEF value at 1.0. MEF is related to the *NSE* metric introduced by Nash & Sutcliffe (1970) that is widely used in hydrological modelling (MEF = [1-NSE]½), but the square-rooting of NSE in MEF suppresses the importance of mismatches of high value events (river run-off, spring bloom of phytoplankton). In a thorough evaluation of NSE metric McCuen *et al.* (2006) advocated for the use of MEF instead of NSE; Zhong & Dutta (2015) applied both Root-Mean-Square-Error (RMSE), NSE and MEF comparing models of operation and maintenance of Light Rail Systems. They concluded that NSE and MEF were superior to RMSE that they had applied earlier.

4 Validation Criteria

As just described we will use three metrices during the validation of the different models. However, applying metrices only makes sense if prior validation criteria have been set. During the review a number of different studies were analyzed and from those studies we also condensed criteria to be applied during the model development. These criteria are reported in Table 4-1, and in this table we also compare to other metrices to allow for a comparison between metrices.

Table 4-1 Use of skill-metrics in evaluating "goodness-of-fit" in coupled aquatic hydrodynamic-ecosystem models. Overview of skill-metrics applied in coupled hydrodynamic-biogeochemical models, and suggested assessment values for individual metrics. Green colours indicate the three metrices used for development of the mechanistic models.

		Azevedo 2014	Allen 2007	Olsen 2016	Haller 2015	Doney 2009	Lehmann 2009	Miladinova 2016	Edwards 2012	Gilboa 2009	Stow 2003	DHI 2017	Assessment values of model skill metrics				References for assessment values [a]
Regression													Excellent	Very good	Good	Poor	
Pearson	r			x		x		x	x	x	x		>0.8	0.8-0.6	< 0.6		"Rooted" r^2
Spearman rank	ρ			x								x	>0.9	0.9-0.6	0.6		Olsen et al. 2016
Coeff determ.	r^2		x	x									>0.65	0.65-0.35	< 0.35		Maréchal 2004
Error metrics																	
Root mean square error	RMSE	x	x	x	x	x	x	x	x		x						
Average error (Bias)	AE			x			x	x	x	x	x						
Average absolute error	AAE	x		x		x					x						
Percent bias	P-bias		x			x			x			x	<10	10-20	20-40	>40	Maréchal 2004, Holt *et al.* 2005, Allen *et al.* 2007
Modelling efficiency	MEF		x	x			x				x	x	>0.8	0.8-0.5	0.5-0.2	< 0.2	
Nash and Sutcliffe 1970	E1	x								x			>0.65	0.65-0.5	0.5-0.2	< 0.2	Legates & McCabe 1999[a], Henriksen *et al.* 2003[a]
Skvar (SDm/SDo)					x												
Index of agreement	d				x												
Cost Function	CF				x				x				0.4	0.4-1	1-2	> 3(5)	Radach & Moll 2006, OSPAR 1998, Holt *et al.* 2005
Parameters																	
T, S			T		T,S	T	T	T,S				T,S					
Satellite data (Chl-a, SST)						x	x	x									
Nutrients		x	x									x					

		Azevedo 2014	Allen 2007	Olsen 2016	Haller 2015	Doney 2009	Lehmann 2009	Miladinova 2016	Edwards 2012	Gilboa 2009	Stow 2003	DHI 2017	Assessment values of model skill metrics				References for assessment values [a)]
Phytoplankton (Chl-a, PP, PC)		C	x			C,P P,P C	C					C,P P					
Total suspended solids		x								x							
Secchi Depth										x		x					
Higher trophic levels				x													

a) References to the different studies are shown below

b) Applied in hydrological models only

5 References

Allen JI, Somerfield PJ & FJ Gilbert (2007) Quantifying uncertainty in high-resolution coupled hydrodynamic-ecosystem models. Journal of Marine Systems 64: 3–14

Azevedo IC, Bordalo AA & P Duarte (2014) Influence of freshwater inflow variability on the Douro estuary primary productivity: A modelling study. Ecological Modelling 272: 1-15

Bennett ND, Croke BFW, Guariso G, Guillaume JHA, Hamilton SH, Jakeman AJ, Marsili-Libelli S, Newham LTH, Norton JP, Perrin C, Pierce SA, Robson B, Seppelt R, Voinov AA, Fath BD & V Andreassian (2013) Characterising performance of environmental models. Environ Model Software 40: 1-20.

DHI (2017a). Hydrodynamic models

DHI (2017b). Biogeochemical model

Doney SC, Lima I, Moore KM, Lindsay K, Behrenfeld MJ, Westberry TK, Mahowald N, Glover DM & T Takahashi (2009) Skill metrics for confronting global upper ocean ecosystem-biogeochemistry models against field and remote sensing data. Journal of Marine Systems 76: 95–112

Edwards KP, Barciela R & M Butenschön (2012) Validation of the NEMO-ERSEM operational ecosystem model for the North-West European Continental Shelf. Ocean Science 8: 983–1000

Gilboa Y, Friedler E & G Gideon (2009) Adapting empirical equations to Lake Kinneret data by using three calibration methods. Ecological Modelling 220: 3291–3300

Haller M, Janssen F, Siddorn J, Petersen W & S Dick (2015) Evaluation of numerical models by Ferry Box and fixed platform in situ data in the southern North Sea. Ocean Sci 11: 879-896

Henriksen HJ, Troldborg L, Nyegaard P, Sonnenborg TO, Refsgaard JC & B Madsen (2003) Methodology for construction, calibration and validation of a national hydrological model for Denmark. Journal of Hydrology 280: 52-71.

Holt J, Allen JI, Proctor R & F Gilbert (2005) Error quantification of a high-resolution coupled hydrodynamic-ecosystem coastal-ocean model: Part 1 model overview and assessment of the hydrodynamics. J Marine Syst. 57: 167–188, doi:10.1016/j.jmarsys.2005.04.008.

Krause P, Boyle DP & F Bäse (2005) Comparison of different efficiency criteria for hydrological model assessment. Advances in Geosciences 5: 89-97.

Legates DR & GJ McCabe (1999) Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic model validation. Water Resources Res. 35(1): 233-241.

Lehmann MK, Fennel K & R He (2009) Statistical validation of a 3-D bio-physical model of the western North Atlantic. Biogeosciences 6: 1961–1974

Maréchal D (2004) PhD Thesis; A soil-based approach to rainfall-runoff modelling in ungauged catchments for England and Wales; Cranfield University at Silsoes UK, 145 p.

McCuen R, Knight Z & A Cutter (2006) Evaluation of the Nash-Sutcliffe Efficiency Index. J Hydrol Eng 11(6): 597-602

Miladinova S, Stips A, Garcia-Gorriz E & DM Moy (2016) Black Sea ecosystem model: setup and validation; EUR 27786; doi: 10.2788/601495. Technical report by the Joint Research Centre; JRC100554.

Moriasi D, Wilson B, Douglas-Mankin K, Arnold J & P Gowda (2012) Hydrologic and water quality models: Use, calibration, and validation. Trans. ASABE 55(4): 1241-1247

Moriasi DN, Gitau MW, Pai N & P Daggupati (2015) Hydrologic and water quality models: Performance measures and evaluation criteria. Trans. ASABE 58: 1763–1785.

Nash JE & JV Sutcliffe (1970) River flow forecasting through conceptual models: Part I. A discussion of principles. Journal of Hydrology 10(3): 282-290.

Olsen E, Fay G, Gaichas S, Gamble R, Lucey S & JS Link (2016) Ecosystem Model Skill Assessment. Yes We Can! PLoS ONE 11(1): 1-24

OSPAR Commission (1998) Report of the Modelling Workshop on Eutrophication Issues. 5–8 November 1996. Den Haag, the Netherlands. OSPAR Report, 86 pp.

Radach G & A Moll (2006) Review of three-dimensional ecological modelling related to the North Sea shelf system. Part II: model validation and data needs. Oceanog. Mar. Biol. 44: 1–60 (An Annual Review).

Stow CA, Roessler C, Borsuk ME, Bowen JD & KH Reckhow (2003) A comparison of estuarine water quality models for TMDL development in the Neuse River Estuary. Journal of Water Resources Planning and Management 129: 307–314.

Stow CA, Jolliff J, McGillicuddy Jr. DJ, Doney SC, Allen JI, Friedrichs MAM, Rose KA & P Wallhead (2009) Skill assessment for coupled biological/physical models of marine systems. J Mar Syst 76(1-2): 4-15

Zhong X & U Dutta (2015) Engaging Nash-Sutcliffe Efficiency and Model Efficiency Factor indicators in selecting and validating effective Light Rail System operation and maintenance cost models. J Traffic and Transportation Engineering 3: 255-265