Available online at www.sciencedirect.com

SCIENCE DIRECT®

Ocean Modelling 11 (2006) 214–223

Ocean Modelling

www.elsevier.com/locate/ocemod

Short Communication

Event-driven model skill assessment

Robert D. Hetland *

Department of Oceanography, Texas A&M University, College Station, TX 77843-3146, United States

Received 29 October 2004; received in revised form 2 December 2004; accepted 2 December 2004
Available online 24 January 2005

Abstract

We expect a numerical simulation to improve upon our first guess; typically some sort of climatology. If a model is not better than this initial estimate of the ocean state, the model provides no new information and serves no purpose. Very often in coastal regions we are interested in predicting events—episodes when the ocean state differs greatly from the climatology. The ability of a model to predict these events will define its usefulness, as these are the deviations from climatology that are unknown before the model is run. An idealized time-series of an event is used to investigate the effects various errors have on the prediction of model skill, a commonly used metric of a model's ability to reproduce observations. It is shown that the choice of climatology is as important in determining model skill as other sources of model error.

1. Introduction

How good is a numerical simulation? There are a variety of methods used to answer this basic question. Most quantitative measures comparing models to observations use some sort of statistical measure. This may be as simple as comparing means and variances of modeled and observed time-series. Commonly, the covariance between modeled and observed time-series is calculated. A more advanced method uses adjoint models to derive a χ^2 goodness-of-fit measure that can be

* Tel.: +1 979 458 0096; fax: +1 979 845 6331.
E-mail address: hetland@tamu.edu

1463-5003/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ocemod.2004.12.001

used to test the hypotheses inherent in choosing model and data errors (Bennett, 2002). Here, we focus on a commonly used metric referred to as model 'skill' (e.g., Bogden et al., 1996). Model skill is, in essence, an estimate of normalized model error variance, and therefore easily calculated. Also, the output is a single number, so that skill can be used to easily and quickly compare a large number of runs.

This paper focuses in particular on defining the skill of a model at reproducing events, defined as abrupt deviations from climatology. In this case, the skill of a model is related to predicting the magnitude and duration of these deviations from climatology. An idealized, analytical model of skill is used to show how the manner in which the model predicts the event (magnitude and timing) relates to the skill. Model errors in predicting these characteristics of the event degrade the skill, as expected. However, the magnitude of background noise and different definitions of the climatology also affect the skill. These additional influences on the skill, separate from the model's true ability to reproduce the event in question, can occasionally have unexpected results on the estimate of the model skill.

The purpose of the paper is to better understand how many different effects come together to influence the value of the calculated model skill. Once the different influences on a skill estimate are understood, the value of the skill may be better interpreted. For example the model skill at predicting an event may be low because of relatively high noise in the data, or it may be artificially high because of highly correlated properties in the model and data, due to tides for example, unrelated to the event in question.

2. Definition of skill

There are many ways to determine model skill, all involve normalization of model error variance. The method of normalization varies from study to study. Oke et al. (2002) use the variance of a reference model to normalize the error variance to test model sensitivity. Holloway and Sou (1996) compare topographic stress parameterizations by calculating skill based on normalized error kinetic energy. The definition of model skill used by Bogden et al. (1996) is also used in this paper:

$$\text{skill} = 1 - \frac{\sum_{i=1}^{N}(d_i - \mathscr{L}[m_i])^2}{\sum_{i=1}^{N}(d_i - c_i)^2}, \tag{1}$$

where d_i are the available measurements, and $\mathscr{L}[m_i]$ is a row vector of the model results in which m_i is transformed by the linear operator $\mathscr{L}$ to match the measurements (see Bennett, 2002), and c_i is a vector of climatological, or background values. The climatology acts as a first guess, and may take a variety of forms. For example, if the measurements are the time it takes an acoustic pulse to travel from the sound source to a receiver, $\mathscr{L}$ will be the appropriate integral of density (which controls sound speed) over the model domain. The final term in the definition of skill can be interpreted the model error variance normalized by the data variance. Thus, a perfect model ($d_i = \mathscr{L}[m_i]$) has a skill of one. If the model simply returns the initial best guess of climatology ($m_i = c_i$), the skill is zero. Note that an energetic model that disagrees with the data may have negative skill.

Model skill is related to the data error term of penalty functionals used in variational data assimilation (Bennett, 2002), where the data error is normalized by a prior estimate of the data error variance. Bogden et al. (1996) use this to create a consistency criterion to estimate the ratio between model error (from open boundaries) and measurement error.

3. Examples

The focus of this study is non-stationary time-series, in particular, an idealized event parameterized as a step function. A step function is chosen because of its relevance to non-stationary events in time-series of coastal and regional processes. Regional numerical simulations generally predict short timescale and small spatial-scale features that are too quick and small for a larger-scale model to resolve. Examples include the response of the ocean to impulsive forcing, such as a storm or a freshet, or a semi-permanent change in ocean state, such as the onset of stratification in the spring. At the same time, available measurements usually only cover a single year or season, so that strong events such as a freshet are sampled once.

Also, measurements near strong frontal regions will measure the values on one side of the front or the other, as the frontal position wafts across the mooring location. In this case, there will be a number of short lived events, and the measured property distribution will be bimodal, representing the properties on either side of the front. The step function also includes the case in which there is a series of events, since the events may be gathered together at the tail end of the time-series; skill does not depend on the order of the data within a time-series. Thus, a step function may mimic a change in state, or a series of impulses, depending on how the time-series is rearranged.

The idealized time-series will be broken down into three components: a step, random noise, and a climatology. Each of these components may vary in amplitude, and the step may occur at different times in the time-series. So, for example, the data time-series will be

$$d(t) = H_d(t) + x_d(t) + c_d, \tag{2}$$

where $x_d(t)$ is a random time-series with a variance of σ_d^2, c_d is the (constant) data climatology, and the step function is defined by

$$H_d(t) = \begin{cases} 0 & : \quad t < (1-\alpha)T, \\ s_d & : \quad t \geqslant \alpha T, \end{cases} \tag{3}$$

where α is the percentage of the duration of the time-series covered by the event, or series of events after the time-series has been rearranged. There is a similar equation for the model time-series. The components of the time-series are shown graphically in Fig. 1. Because of their simple structure, these time-series may be put into the skill equation and calculated analytically. Below, a few simple, common cases have been calculated, starting with a simple comparison of signal to noise.

3.1. Signal to noise

The first question to be addressed is how random noise will affect an otherwise perfectly reproduced event. Let α be the fraction of the time-series covered by the event, the second half of the

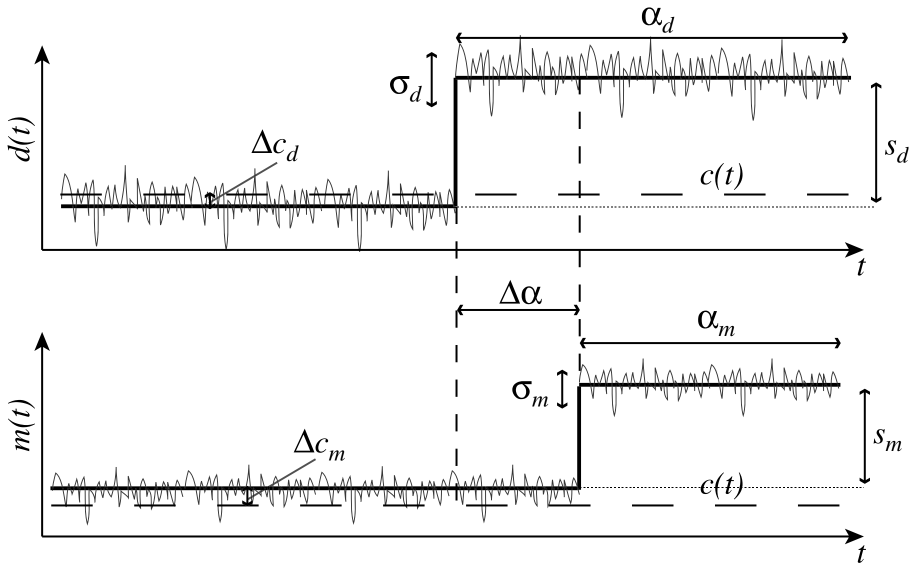

Fig. 1. An illustration of the variables in the idealized step model/data comparison. Variables are defined in the text. Three time-series are indicated: $m(t)$ is the model, $d(t)$ the data, and $c(t)$ the climatology. Subscripts d and m indicate data and model time-series, respectively. Each time-series contains an event, with a step-like increase of s, the standard deviation of the noise within each time-series is σ. Δc is the differences in climatology between base state in the data and in the model time-series. The fraction of the time-series covered by the event is α. The fraction of the time-series when an event is erroneously predicted or erroneously absent is $\Delta\alpha$.

step that deviates from climatology. The standard deviation of uncorrelated noise in both the model and observations are identical, so that $\sigma = \sigma_m = \sigma_d$. The magnitude of the event in both the model and observations is $s = s_m = s_d$. In the case with no step, $s = 0$, the skill is -1. A non-zero step increases the skill, defined for this case as

$$\text{skill} = 1 - \frac{2}{\alpha\left(\frac{s}{\sigma}\right)^2 + 1}. \tag{4}$$

Note that in the limit of infinite signal to noise, the skill is that of a perfect model. The skill is zero when $\alpha s^2 = \sigma^2$, requiring that the variance associated with the step, αs^2, be larger than the variance of the background noise for positive skill values to be achieved. Note that the variance of the event depends on both the magnitude of the step and the duration of the event. Thus, we have the intuitive result that an increased signal-to-noise ratio increases model skill.

3.2. Differences in noise amplitude

Often, model variance is lower than measured variance. Consider again a perfectly modeled step, but now with uncorrelated noise of different magnitudes in the model and observations. The skill estimate in this case is

$$\text{skill} = 1 - \frac{\left(\frac{\sigma_m}{\sigma_d}\right)^2 + 1}{\alpha\left(\frac{s}{\sigma_d}\right)^2 + 1}, \tag{5}$$

which shows that a decrease in model noise variance increases the skill more than an equivalent decrease in data noise variance. Because of this, skill estimates using a model with low noise variance have a higher model skill than models with noise variance that match the data noise variance. Thus, it is possible that lower model skill may be due to higher noise variance rather than a degradation in event prediction.

3.3. Differences in amplitude

What if the model predicts the timing of the event, but predicts a different amplitude? Assume that the noise within the model and data are the same ($\sigma = \sigma_d = \sigma_m$), as is the duration of the modeled and predicted event ($\alpha = \alpha_d = \alpha_m$). In this case, the skill is

$$\text{skill} = 1 - \frac{\alpha(\Delta s)^2 + 2\sigma^2}{\alpha s_d^2 + \sigma^2}, \tag{6}$$

where $\Delta s = |s_d - s_m|$. In the limit of low noise ($s^2 \gg \sigma^2$ and $\Delta s^2 \gg \sigma^2$), the skill reduces to

$$\text{skill} = 1 - \left(\frac{\Delta s}{s_d}\right)^2. \tag{7}$$

The skill decreases with increasing error in amplitude prediction, an effect enhanced by longer event duration and decreased with a larger signal-to-noise ratio. An error in 10% in the predicted event amplitude over half of the time-series, in an otherwise perfect simulation, will result in a skill of 0.99. Thus, the skill may be relatively insensitive to the ability of the model to predict the event amplitude.

3.4. Differences in event duration

What if the model captures the strength of the event, but the timing is wrong? This effect may be included in our simple time-series mode by a difference in the fraction of the time-series covered by the event. Note that the error will be the same for both a leading and lagging prediction. Separate event fractions for the data and model, α_d and α_m respectively, are now required. The fraction of the time-series that does not overlap, where the model fails to predict either the presence or absence of the event, is $\Delta\alpha$ (not necessarily $|\alpha_d - \alpha_m|$). The skill in this case is

$$\text{skill} = 1 - \frac{\Delta\alpha s^2 + 2\sigma^2}{\alpha_d s^2 + \sigma^2}, \tag{8}$$

where it has again been assumed that $s = s_m = s_d$ and $\sigma = \sigma_d = \sigma_m$. Again assuming a large signal to noise ratio ($\alpha s^2 \gg \sigma^2$), this equation reduces to

$$\text{skill} = 1 - \frac{\Delta\alpha}{\alpha_d}. \tag{9}$$

Comparing Eqs. (7) and (9), it is clear that the skill is relatively more sensitive to errors in predicting event duration than to errors in predicting event amplitude (assuming, of course, $\Delta s < s_d$ and $\Delta\alpha < \alpha_d$).

3.5. Offset in climatology

Using our present definition of skill, the background, climatological time-series is removed from both model and observations before the skill is calculated. It was assumed in the previous examples that the initial portion of the time-series was identical to the climatology. Here we examine the effect of an offset in both measured and modeled time-series. Now assume that the measured and modeled time-series are identical ($s = s_m = s_d$ and $\sigma = \sigma_m = \sigma_d$), except for an offset of Δc_d and Δc_m, respectively. The skill estimate becomes

$$\text{skill} = 1 - \frac{(\Delta c_d - \Delta c_m)^2 + 2\sigma^2}{\sigma^2 + \alpha(s + \Delta c_d)^2 + (1-\alpha)\Delta c_d^2}. \tag{10}$$

The denominator is smallest, and the skill lowest, when $c_d = \alpha_s$, that is, the climatology is the mean of the time-series. If the offset in both time-series, $\Delta c_d - \Delta c_m$, is held constant, deviations in the magnitude of c_d away from αs improve model skill. Thus, large differences between the climatology and the mean of the data may artificially inflate the skill estimate. On the other hand, if the model accurately predicts large, real deviations from the climatology, an enhanced skill estimate may be warranted. Only constant values of climatology are considered here, for the sake of mathematical simplicity, but generally, any functional or statistical form of the climatology can be considered.

3.6. Coherent signal superimposed

Often both the model and observed time-series have a high-frequency, coherent signal in common, like tides. Suppose both time-series are divided into noise, with a variance of σ^2_{noise} and no covariance between the two time-series, and a coherent signal with a variance and covariance of $\sigma^2_{\text{coherent}}$. Assuming no offset in climatology, and a perfect simulation of the event portion of the time-series, the skill is

$$\text{skill} = 1 - 2\frac{\sigma^2_{\text{noise}}}{\sigma^2_{\text{noise}} + \sigma^2_{\text{coherent}} + \alpha s^2}. \tag{11}$$

Skill increases with increasing variance of the coherent signal. If the variance of the coherent signal, $\sigma^2_{\text{coherent}}$, is larger than the event variance, αs^2, the ability of the model to reproduce the event will have little influence on the overall model skill.

4. Skill estimates from oceanic measurements

An example of a time-series with a statistically significant event, shown in Fig. 2, demonstrates how different features of the observed and simulated time-series combine to form the skill

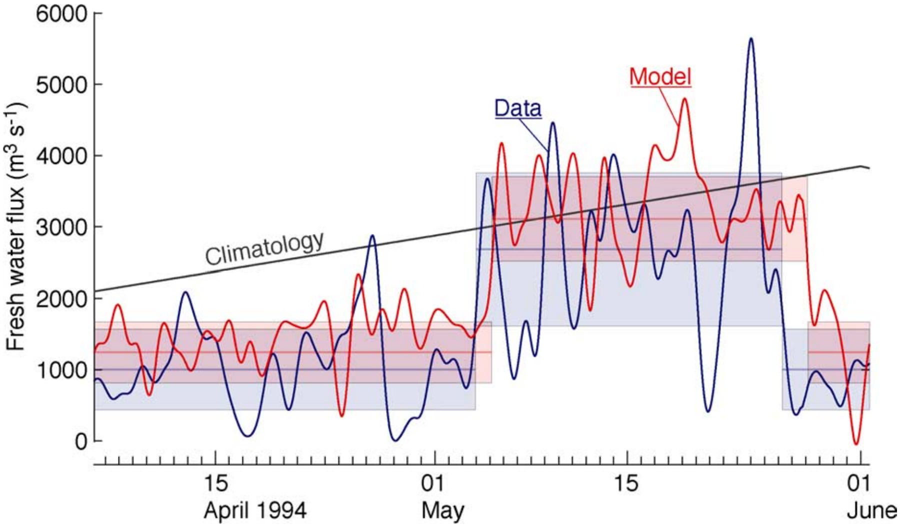

Fig. 2. A time-series of fresh water flux carried within the eastern Maine coastal current, estimated from observations and a numerical model shows a statistically significant event between early May and late May. Thick solid lines show the subtidal (33 hour low pass filtered) fresh water flux. Lighter solid horizontal lines show the means, and shading shows the standard deviation, during distinct periods of high and low fresh water flux for each time series in matching colours. Climatology is shown as a solid gray line.

estimate. The time-series is an estimate of net along-shore fresh water flux; the details of the calculation are described in Hetland and Signell (accepted for publication), but are not important for the discussion here. Generally, the model used is a high-resolution (2–3 km grid spacing) coastal ocean model of the Gulf of Maine. The model is locally forced with realistic winds and fresh water fluxes for rivers, and the coastal current structure is resolved by the model grid. Boundary forcing includes tides and nudging mean flow and tracer fields to climatological values. Tides have been filtered out of this time-series, since the purpose of the study was to examine the subtidal circulation, and including the tides would artificially inflate the skill.

The primary question that Hetland and Signell sought to answer was: Is the model good enough to use as a foundation for regional harmful algal bloom prediction? This immediately begs the question of what is meant by 'good enough.' Because harmful algae are carried by the buoyancy-driven coastal current system, accurate prediction of this system is essential, and a metric was devised to measure this feature in the model the fresh water flux. It was clear which simulations were best by comparing skill values. However, it was not as clear if even the best simulation was good enough. In order to determine this, the meaning of the skill value must be interpreted.

The model and data time-series (Fig. 2) both show an event that occurs in early to late May. Event durations were chosen by finding the time-frames that created the most distinct time-series, the largest gap between the means plus or minus the standard deviations. Both time-series have three distinct sections. The first and last sections are assumed to be statistically identical, with the same means and same standard deviations. The middle section, the event in these time-series, has a higher mean and standard deviation than the other sections. These and other time-series statistics are shown in Table 1.

Table 1
Statistical properties are given for the base and event sections of the model and data fresh water flux time-series shown in Fig. 2

	'Base' (first and last sections)	'Event' (middle section)
Model		
Mean ($m^3 s^{-1}$)	1244.2	3109.2
Standard deviation ($m^3 s^{-1}$)	427.7	592.8
Duration (%)	59	41
Data		
Mean ($m^3 s^{-1}$)	1002.8	2685.8
Standard deviation ($m^3 s^{-1}$)	564.4	1071.5
Duration (%)	60	40

The estimated model skill may be compared to the analytical skill estimates from Section 3. There is a 95% overlap in the time-series, so that $\Delta\alpha = 0.05$. This results in a skill of 0.88, given a high signal-to-noise and perfect reproduction of the step amplitude (Eq. (9)). The signal-to-noise ratio—the ratio of the difference in the means to the mean standard deviation excluding the step—is 2.2 in the data and 3.8 in the model. This leads to a skill of 0.57, everything else assumed perfect (Eq. (5)). The difference in predicted event amplitude is about 10%, resulting in a skill of 0.99, all other detrimental factors excluded (Eq. (7)). It is apparent that the signal to noise ratio is the largest factor in degrading the skill when the climatology is assumed to be equal to the base value of the time-series. Filtering the time-series results in substantial improvements in the skill.

The climatology shown in Fig. 2, calculated from the same fields used to initialize the model and provide boundary conditions, shows a linear trend that roughly reproduces the increase seen in the higher resolution time-series. The skill of the model using this climatology is 0.56. Using the observed mean of the time-series base, the skill reduces to 0.29 (Eq. (10)). This is approximately the skill we would predict by summing the effects described in the preceding paragraph. A minimum in skill of 0.05 is found using a constant climatological value of 1669. Thus, referencing the time-series to the actual climatology, as opposed to the mean of the observed fresh water flux, produces a much enhanced skill value. This may be justified, given that the model was initiated with the climatology, and maintained climatological conditions along the boundaries throughout the simulation. This means that differences between the climatology and the simulation are due to local wind and river forcing in the model. Thus, correctly reproducing deviations in the climatology from observed time-series means that the model has skill in simulating these locally forced processes, and the model skill should take this into account.

5. Discussion and conclusions

Simple analytical models of skill have been derived for an idealized time-series containing an event (or events) which deviate from climatology. These analytical models are useful in determining the factors that determine the model skill. Understanding the factors that contribute to skill estimates allow one to ascertain whether or not a particular skill value is good or not, given the circumstances of the simulation. Also, understanding the factors that degrade skill can be used to direct model improvements.

Skill estimates typically vary in response to changing time-series statistics as expected. The results from Section 3 shows

a. increased signal-to-noise increases skill,
b. decreased model noise variance increases skill,
c. differences in actual an predicted event amplitude affect skill, but this effect may be be small,
d. differences in predicted event duration, or a mistimed event, decrease skill, and
e. offsets in climatology different than the mean of the data can increase.

Individually, none of these results are particularly surprising. However, by quantitatively comparing the relative effects of each of these influences, the dominant influences on the skill estimate may be determined. For example, in the case study discussed in Section 4, the definition of climatology had the largest effect on the skill estimate. Using simple means of the time-series as reference levels results in the lowest possible skill. This may be desired in many situations, as the most stringent test of the model's ability, however, there may be times when using different definitions of the climatology is desirable and justified. Differences in event amplitude had nearly no effect on the skill estimate, so the definition of skill used in this paper would not be useful in evaluating simulations where amplitude prediction is critical.

It is not necessarily true that skill calculations for all applications will be most sensitive to the climatology. All of the influences on skill scores must be estimated for each case to determine which dominate. Thus, the interpretation of skill scores is subjective and application dependent.

It would be a simple matter to extend this theory to include more complicated climatologies, events, and additional processes. For example, examining model skill at predicting deviations from an annual cycle. Creating these models forces one to consider the different processes acting within the time-series, and how these processes interact to affect the model's ability to reproduce observations. If a particular feature of a model is determined to be critical, special metrics (e.g., a modified skill score equation) should be created that are sensitive to these critical processes and insensitive to less critical aspects of the model.

Acknowledgements

I would like to thank Steve Baum, Steve DiMarco, Ann Jochens, Worth Nowlin and Bob Ried for reviewing drafts of this manuscript. This work was supported by the Office of Naval Research through the Distributed Marine Environmental Forecast System program (N000140010886).

References

Bennett, A.F. Inverse Modeling of the Ocean and Atmosphere. Cambridge, 2002.

Bogden, P.S., Malanotte-Rizzoli, P., Signell, R.P., 1996. Open-ocean boundary conditions from interior data: Local and remote forcing of Massachusetts Bay. J. Geophys. Res. 101, 6487–6500.

Hetland, R.D., Signell, R.P. Modelling coastal current transport in the Gulf of Maine. Deep-Sea Res. II, accepted for publication.

Holloway, G., Sou, T., 1996. Measuring skill of a topographic stress parameterization in a large-scale ocean model. J. Phys. Oceanogr. 26, 1088–1092.

Oke, P.R., Allen, J.S., Miller, R.N., Egbert, G.D., Austin, J.A., Barth, J.A., Boyd, T.J., Kosro, P.M., Levine, M.D., 2002. A modeling study of the three-dimensional continental shelf circulation off Oregon. Part I: Model–data comparisons. J. Phys. Oceanogr. 32, 1360–1382.