geosciences

MDPI

Long-Term Changes, Inter-Annual, and Monthly Variability of Sea Level at the Coasts of the Spanish Mediterranean and the Gulf of Cádiz

Manuel Vargas-Yáñez[1,*], Elena Tel[2], Francina Moya[1], Enrique Ballesteros[1] and Mari Carmen García-Martínez[1]

[1] Centro Oceanográfico de Málaga, Instituto Español de Oceanografía (Consejo Superior de Investigaciones Científicas, CSIC), Puerto Pesquero s/n, Fuengirola, 29640 Málaga, Spain; francina.moya@ieo.es (F.M.); enrique.ballesteros@ieo.es (E.B.); mcarmen.garcia@ieo.es (M.C.G.-M.)

[2] ServiciosCentrales, InstitutoEspañol de Oceanografía, (Consejo Superior de Investigaciones Científicas, CSIC) C. del Corazón de María, 8, 28002 Madrid, Spain; elena.tel@ieo.es

* Correspondence: manolo.vargas@ieo.es

Filling gaps using multiple linear regression.

Let y be a variable for which we have n observations. Each observation is denoted by y_i, $i=1,...n$.

Let $\{x_k\}$be a set of m variables. Hereafter y will be considered as the dependent variable, and x_k will be the independent variables or predictors. We also have n observations of the predictors or independent variables corresponding to the n observations of the dependent variable. That is: we have $x_{i,k}$ with $i=1,...n$, $k=1,...m$

In the present case, y will be the sea level measured at a certain location, and x_k will be the sea level measured at m nearby locations. Each of the n observations of the dependent and independent variables correspond to the same time.

Initially we consider that there is a linear relationship between the dependent variable (the sea level time series that we want to reconstruct), and the independent variables or predictors (see level at the close tide gauges). This can be expressed as:

$$y_i = \beta_0 + \sum_{k=1}^{m} \beta_k x_{i,k} + \varepsilon_i \text{ , i=1,...n}$$

β_0 is the interception at the origin and ε represents the part of y that is not reproduced by the linear model.

This equation can be expressed in matrix form:

$$y=X\beta+\varepsilon \qquad \text{(S.1)}$$

y is a column vector with the n observations of the dependent variable.

X is a$nx(m+1)$ matrix, where all the elements of the first column are equal to 1, and the other m columns are the observations of the predictors x_k.

$\beta$is a $(m+1)x1$ column vector with the true (unknown) coefficients that relate the sea level time series (y) with the sea level at the nearby tide gauges x_k. Notice that this relation could not exist, and the true value of the coefficients, could be zero.

$\varepsilon$is a n x 1 column vector with the part of y not explained by the linear model and will be named as the residuals hereafter.

These coefficients will be estimated minimizing the sum of the squares of the residuals (least squares fit). If b represents the estimation of β, we have to minimize the expression:

$$(y-Xb)^T (y-Xb)$$

where T denotes transpose.

Deriving respect to b, we obtain the set of $m + 1$ equations:

$$\left(XX^T\right)b = X^T y \text{ (S.2)}$$

And the estimated coefficients are:

$$b = \left(XX^T\right)^{-1} X^T y \text{ (S.3)}$$

The residuals can be estimated as:

$$\hat{\varepsilon} = y - Xb \text{ (S.4)}$$

Even in the case that the sea level y was not related with the sea level at the nearby locations used in the regression, the expression (S.3)could yield values of b different from zero. Therefore we should determine whether the linear relation obtained is significant or not from a statistical point of view.

Then we make the null hypothesis (H_0):

$$H_0\text{: all } \beta_k,\ k=1,\ldots m \text{ are equal to zero}$$

Then we test this hypothesis. To do so, it can be shown that the statistic:

$$F = \frac{b^T\left(X^T X\right)b \,/\, m}{\hat{\varepsilon}^T\varepsilon \,/\, n-m-1} \text{ (S.5)}$$

follows adistribution F-Fisher with m and n-m-1 degrees of freedom.

Notice that in this case the b and the X in the numerator only contain the values and columns corresponding to the predictors (not the column of 1).

Once the value F has been calculated, we can consider the following question: Considering that the null hypothesis is true (the β_k are zero) which is the probability of obtaining a value of F as large as the one we have obtained. This question can be answered using the cumulative probability function $F_{m,\ n\text{-}m\text{-}1}$. If the probability of obtaining the F value is lower than 0.05 (another threshold could be used), we can say that it is very unlikely that this value has been obtained by chance, and we reject the null hypothesis. Therefore, we accept that the β_k are different from zero and thatthe linear regression is significant at the 0.05 significance level (we insist that other significance level could be used).

Expression (S.5) shows the ratio between the variance of y expressed by the linear model and the variance of the residuals, that is, the part not explained by the linear model. The interpretation of this result is that, if F is very large, the linear model explains a large fraction of the variance of y, much larger that the part out of the model. Then, we cannot accept that the linear relation does not exist.

We can also estimate the percentage or fraction of the variance of y, explained by the model. If $\bar{y}$ is the sample mean of the observations:

$$R^2 = \frac{\left(Xb - \bar{y}\right)^T\left(Xb - \bar{y}\right)}{\left(y - \bar{y}\right)^T\left(y - \bar{y}\right)}$$

And the square root of this expression is R, the multiple correlation coefficient.

Even in the case that we reject the null hypothesis and we accept that the linear regression is statistically significant, it could be true that the contribution of some of the predictors (x_k) were not significant. In other words, the variance explained by the linear model, $b^T(XX^T)b$ could be large, but the contribution of some of the predictors could be omitted from the model. We have rejected the null hypothesis: All the $\beta_k=0$. But still, some of them could be zero.

In other to test the significance of the contribution of each single predictor, we estimate the statstic F-partial.

For instance, let us consider that we want to test the significance of the contribution of the x_m independent variable. First we fit a linear model without including x_K:

$$y_i = \beta_0 + \sum_{k=1}^{m-1} \beta_k x_{i,k} + \varepsilon_i$$

Then the sum of squares explained by this model (without x_m) can be calculated:

$$S1 = b^T \left(XX^T \right) b$$

Then, we include x_m in the model and estimate the coefficients (which in general will not be the same) and estimate the new sum of squares estimated by the model (including x_m): S2.

The statistic:

$$F_{partial} = \frac{S2 - S1}{\hat{\varepsilon}^T \hat{\varepsilon} \big/ n-m-1}$$

is distributed as a F-Fisher with 1 and n-m-1 degrees of freedom.

Once again, the probability of obtaining such F-partial value can be calculated. If this probability is lower than a certain threshold, we accept that the contribution of x_m is significant and we retain it in the model . Otherwise, this predictor should be removed from the model.

Once the coefficients of the regression have been determined, they can be used to estimate the value of y for that times where the predictors are known and the dependent variable is unknown.

Table S1. shows the tide-gauges used for the analysis of long-term changes, the tide gauges used for the linear multiple regression,the F values for the regression, its probability (p-value), the multiple correlation coefficients, and the coefficients of the linear regression.

Cádiz II	**Ceuta Málaga Tarifa**	**F = 42.51 p ~0**	**R = 0.69**	**b_0= 88±9 b_1=0.23±0.13 b_2=0.41±0.10 b_3=0.34±0.14**
Algeciras	Cádiz III Ceuta Málaga Tarifa	F =213.47 p ~0	R=0.84	b_0= 18±3 b_1=0.04±0.04 b_2=0.24±0.06 b_3=0.37±0.08 b_4=0.19±0.06
Tarifa	Algeciras Cádiz III Málaga	F=124.23 p~0	R=0.7	b_0= -7±5 b_1=0.39±0.14 b_2=-0.05±0.05 b_3=0.48±0.12
Ceuta	Málaga	F=269.30 p~0	R=0.54	b_0=-3±3 b_1=0.45±0.05
Málaga	Cádiz III Ceuta Tarifa	F=418.92 p~0	R=0.84	b_0=-2±3 b_1=0.18±0.04 b_2=0.14±0.08 b_3=0.61±0.06
Alicante_out	Alicante_in	F=1309 p~0	R=0.84	b_0=19±2 b_1=0.87±0.05
L'Estartit	Barcelona	F=1582 p~0	R=0.92	b_0=8±3 b_1=0.80±0.04

It should be noticed that the tide gauges shown in the second column of table S1 are those used when data were available. Some periods of time had to be reconstructed using a lower number of predictors.

Forward stepwise regression. Influence of meteorological factors and steric level.

In this case the dependent variables are the monthly time series of reconstructed sea level. First, the average seasonal or annual cycle was subtracted. The procedure was the following. First, a mean value was calculated for each month of the year using the complete time series. For instance, the time series of sea level at Málaga has 65 years that extend from 1944 to 2018. The mean value of the 65 months of January was calculated. Similarly, the mean value for the 65 months of February was calculated, etc. In this way we obtain 12 values that represent the average seasonal or annual cycle. Then this cycle was subtracted to the original time series. Then a straight line was fitted to the de-seasoned time series and was also subtracted. This procedure was also applied to the monthly time series of atmospheric pressure (P), U and V components of the wind, and the thermosteric (η_T) and halosteric variability of sea level (η_H).

For each tide gauge, the variability of sea level is caused by changes in the atmospheric variables, which can induce local redistributions of mass, and changes in the density of sea water. We could propose the following linear model to explain the variability of sea level:

$$\eta = \beta_0 + \beta_1 P + \beta_2 U + \beta_3 V + \beta_4 \eta_T + \beta_5 \eta_5 + \varepsilon \quad (1)$$

Nevertheless, we have considered the following approach. It was not assumed a priori which variables have a significant influence on the sea level variability. P, U, V, η_T, η_H were considered as candidate predictors. We calculated the correlation between the sea level (η), and each of the possible predictors. That predictor which the highest correlation was considered as the first candidate to be included in the linear model. Just to fix ideas,

let us consider that P was the variable with the highest correlation. Then, the proposed model is: $\eta = \beta_0 + \beta_1 P + \varepsilon$. As always, ε is the part of η not explained by the model. We test the significance of this model at a certain level of significance (in the present work we have used the level 0.05). If the regression is not significant, the procedure is over, and the model selected is simply: $\eta = \beta_0 + \varepsilon$

On the other hand, if the regression is significant, we accept the proposed model $\eta = \beta_0 + \beta_1 P + \varepsilon$ and we consider the possibility of including a new predictor in the model. The following step is to regress all the other candidate predictors (U, V, η_T, η_H) on P. We estimate the residuals for these regressions, that is, we obtain the part of the other predictors not explained by P, and we calculate the correlation between these residuals, and the part of η not explained by P (residuals of the initial linear model). The variable with the highest correlation is considered as a candidate to be included in the linear model. Let us suppose that such variable is U, then the new candidate model is: $\eta = \beta_0 + \beta_1 P + \beta_2 U + \varepsilon$

We test the significance of this new model, and we also test the F-partial for each of the predictors in the model. If the new regression is significant, and the contribution of all the predictors (F-partial) is significant, then we accept the new model $\eta = \beta_0 + \beta_1 P + \beta_2 U + \varepsilon$. If the regression is not significant, the procedure is over and the selected model is $\eta = \beta_0 + \beta_1 P + \varepsilon$. Otherwise, we test the possibility of including a new predictor. It is important to notice that at any step, a predictor that had been included in a previous step, can be excluded if its F-partial becomes not significant when including some new predictors. The significance level to include a new predictor should be lower than the significance level used to get a predictor out of the model. In the present work we have followed the usual criterion of 0.05 to get in, and 0.10 to get out.

Table 4 in the main text of this work, shows the models selected by the forward step-wise regression for each of the sea level time series and for each period of time.