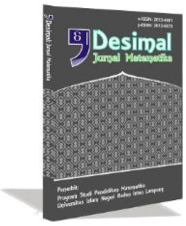
Contents lists available at DJM

DESIMAL: JURNAL MATEMATIKA

p-ISSN: 2613-9073 (print), e-ISSN: 2613-9081 (online), DOI 10.24042/djm

http://ejournal.radenintan.ac.id/index.php/desimal/index



El nino index prediction model using quantile mapping approach on sea surface temperature data

Sri Nurdiati*, Elis Khatizah, Mohamad Khoirun Najib, Linda Leni Fatmawati

IPB University, Indonesia

ARTICLE INFO

Article History

Received	*: 25-11-2020*
Revised	*: 08-03-2021*
Accepted	*: 23-03-2021*
Published	*: 27-03-2021*

Keywords:
ECMWF, El Nino Index, Quantile Mapping, Statistical Bias Correction.

*Correspondence: E-mail: nurdiati@apps.ipb.ac.id

Doi: 10.24042/djm.v4i1.7595

ABSTRACT

El Nino is a global climate phenomenon caused by the warming of sea surface temperatures in the eastern Pacific Ocean. El Nino has a powerful effect on the intensity of rainfall in several areas in Indonesia. El Nino impacts can be minimized by predicting the El Nino index from the sea surface temperature in the Nino 3.4 area. Therefore, many researchers have tried to predict sea surface temperature, and many prediction data are available, one of which is ECMWF. But, in reality, the ECMWF data still contains systematic errors or bias towards the observations. Consequently, El Nino predictions using ECMWF data are less accurate. For that reason, this study aims to correct the ECMWF data in the Nino 3.4 area using statistical bias correction with a quantile mapping approach. This method uses ECMWF data from 1983-2012 as training data and 2013-2018 as testing data. For this case, the results showed that 60% of El Nino's predictions on the testing data had improved the mean value. Also, all of El Nino's predictions on the testing data have improved the standard deviation value. Moreover, data testing's expected error can be corrected for all months in the 1st to 4th lead times. But, in the 5th to 7th lead times, only November-June can be corrected.

http://ejournal.radenintan.ac.id/index.php/desimal/index

INTRODUCTION

Indonesia is a country with a lot of natural resources. But, the environmental damage in Indonesia is relatively high. One of them is caused by climate change. Climate change is already having visible effects on the world. The Earth is warming, rainfall patterns are changing, and sea levels are rising. These changes can increase the risk of floods, droughts, heatwaves, and fires.

Indonesia's climate conditions are influenced by conditions of sea surface temperature, both in the Pacific and Indian Ocean (including local SST in Indonesian seas) (Aldrian & Susanto, 2003). The climatic conditions in the Pacific Ocean are known as El Nino Southern Oscillation (ENSO), and the climatic conditions in the Indian Ocean are known as Indian Ocean Dipole (IOD).

El Nino is an ocean-atmosphere phenomenon on a global scale (Philander, 1989). El Nino occurs when there is a positive difference between observed sea surface temperatures and normal conditions in the equatorial Pacific Ocean. This condition recurs every 3-8 years (Nabilah et al., 2017). El Nino is indicated by the sea surface temperature in the Pacific Ocean, and the difference in air pressure between Darwin and Tahiti increases periodically (Taufik & Marnita, 2004). During El Nino, Indonesia's winds from the Pacific Ocean contain less water vapor, so Indonesia's dry season is more extended (Hannachi, 2004). Low rainfall intensity and long dry season are the direct impacts of El Nino which can trigger various problems.

El Nino's impact can be minimized by predicting the El Nino index using sea surface temperature in the Nino 3.4 area. Many researchers have studied to build prediction models for sea surface temperature. Therefore, many sea surface temperature prediction data are released by the European Centre for Medium-Range Weather Forecasts (ECMWF). But the ECMWF data still contains systematic errors or bias towards its observations. These biases can significantly impact seasonal forecasts and future climate predictions (Shonk et al., 2019).

Consequently, El Nino predictions using ECMWF data are less accurate. So, a method is needed to correct the bias of the sea surface temperature forecast data from the ECMWF. One method that can be used is a statistical bias correction (Misnawati et al., 2018).

Many scientists have studied statistical bias correction methods, such as Piani et al., who designed and applied bias corrections to the output of the DMI-Hirham daily climate model over Europe, resulting in the same distribution as the observation model (Piani et al., 2010). Lealdi et al. used the statistical bias correction method to see the relationship between the ECMWF rainfall data and the BMKG's observation model in the 1996-2015 period for cases on the island of Bali (Lealdi et al., 2018). Dasanto et al. use rainfall data in the watershed (DAS) Citarum Hulu with quantile mapping approach statistical bias correction (Dasanto et al., 2014) and many others (Ayugi et al., 2020; Katiraie-Boroujerdy et al., 2020; Passow & Donner, 2020; Pastén-Zapata et al., 2020).

This study uses the Nino 3.4 area as the observation area for the El Nino index. The Nino 3.4 anomalies may be thought of as representing the average equatorial SSTs across the Pacific from about the dateline to the South American coast (K. Trenberth & National Center for Atmospheric Research Staff, 2020). With the assumption that the HadISST data is observational data representing the actual sea surface temperature Physical Sciences Laboratory NOAA uses the HadISST data to calculate the El Nino index (Rayner et al., 2003). This study uses statistical bias correction with a quantile mapping approach to see the relationship between the ECMWF and the HadISST data from 1983 to 2012 in the Nino 3.4 area. The results will be used to correct the ECMWF sea surface temperature data from 2013 to 2018 and calculate the Nino index prediction model. After that, the Nino index prediction model is evaluated against the actual Nino index. The Nino index prediction model using corrected ECMWF data is expected to have better accuracy than using ECMWF data before correction.

METHOD

Sources and Types of Data

This study's historical sea surface temperature data is the global sea surface temperature data of The Hadley Center Global Sea Ice and Sea Surface Temperature (HadISST) from 1983 to 2018, which can be downloaded on Met

Office Hadley Center official website http://www.metoffice.gov.uk/hadobs/hadisst/data/HadISST_sst.nc. This global HadISST data has the NetCDF (Network Common Data Form) format, and this data is sea surface temperature data that is released every month with a data dimension size of 360 (longitude) × 180 (latitude) × 1798 (month).

The following data is the ECMWF (European Centre for Medium-Range Weather Forecasts) forecast data from 1983-2018, which can be accessed through the Center of Development and Research of the Meteorology, Climatology and Geophysics Agency (BMKG) of Indonesia. The data format is NetCDF. ECMWF data is predictive data for sea surface temperature in the Nino 3.4 area, which has dimensions of 205 (longitude) × 45 (latitude) × 25 (ensemble) × 216 (lead time) for each month from 1983-2018.

Statistical Bias Correction

Statistical bias correction is a technique of connecting a relationship between observed and predicted data to form a transfer function ($y = f(x)$). This function combines the values of the quantile of cumulative distribution function (CDF) of observations and predictions in the form of an equation

$$cdf_{obs}(y) = cdf_{pred}(x) \qquad (1)$$

The relationship between observational and predictive data can be linear, exponential, or polynomial regression equations (Dasanto et al., 2014; Misnawati et al., 2018). The method that connects the observed and predicted data are called statistical downscaling, quantile mapping, histogram equalizing or statistical bias correction (Piani et al., 2010). Misnawati et al. said the first step in statistical bias correction using a quantile mapping approach is to identify the distribution and probability density function (Misnawati et al., 2018). The second step is to compute the cumulative distribution by integrating the probability density function. The third step is to create a transfer function between the observed and predicted data's cumulative distribution.

Research Steps

The data obtained from HadISST and ECMWF are still in the NetCDF format. Therefore, a data extraction process is needed so that the data is ready to be processed. The data extraction process for HadISST data includes cutting time intervals, cutting Nino 3.4 domains, and compiling data matrices. Data is stored as training data from 1983-2012 and testing data from 2013-2018. On the other hand, ECMWF data were obtained between 1983-2018 around the Nino 3.4 area so that the data extraction process only included includes adjusting the time interval and domain of Nino 3.4 and compiling a data matrix, namely training data from 1983-2012 and testing data from 2013-2018 in 1st to 7th data lead times.

The second step is to identify the distribution of data. The distribution identification is executed on the HadISST data and the ECMWF data extracted for each month and the lead times. The distributions were identified using several parametric distributions as follows.

Type-1 Extreme Value (EV) distribution also known as Gumbel distribution(E.J. Gumbel, 1941) for minima:

$$f(x \mid \mu, \sigma) = \sigma^{-1} e^{z} \exp(-e^{z}) \qquad (2)$$

where $z = (x - \mu)/\sigma$.

Generalized Extreme Value (GEV) distribution (Hosking et al., 1985):

$$f(x \mid k, \mu, \sigma) = \frac{1}{\sigma} e^{-z^{-\frac{1}{k}}} z^{-1-\frac{1}{k}} \qquad (3)$$

where $z = 1 + k(x - \mu)/\sigma$.

Logistic (LOG) distribution (Decani & Stine, 1986):

$$f(x \mid \mu, \sigma) = e^{z}/[\sigma(1 + e^{z})^{2}] \qquad (4)$$

where $z = (x - \mu)/\sigma$.

Normal (NOR) distribution or Gaussian distribution (Hald, 1949):

$$f(x \mid \mu, \sigma) = \left(\sigma\sqrt{2\pi}\right)^{-1} e^{-\frac{1}{2}z^2} \qquad (5)$$

where $z = (x - \mu)/\sigma$.

Exponential (EXP) distribution (Walpole, 1990):

$$f(x \mid \mu) = \mu^{-1} \exp(-x\mu^{-1}) \qquad (6)$$

Gamma (GAM) distribution (Kollu et al., 2012):

$$f(x \mid \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-\frac{x}{\beta}} \qquad (7)$$

Inverse Gaussian (ING) distribution (Chhikara & Folks, 1977):

$$f(x \mid \mu, \lambda) = \sqrt{\frac{\lambda}{2\mu^2 x}}\, e^{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}} \qquad (8)$$

Log-logistic (LL) distribution

$$f(x \mid \mu, \sigma) = \frac{1}{x\sigma}\frac{e^z}{(1+e^z)^2} \qquad (9)$$

where $z = (\log x - \mu)/\sigma$.

Log-normal (LN) distribution (Kollu et al., 2012):

$$f(x \mid \mu, \sigma) = \left(x\sigma\sqrt{2\pi}\right)^{-1} e^{-\frac{1}{2}z^2} \qquad (10)$$

where $z = (\ln x - \mu)/\sigma$.

Weibull (WB) distribution (Kollu et al., 2012):

$$f(x \mid a, b) = \frac{b}{a}\left(\frac{x}{a}\right)^{b-1} e^{-\left(\frac{x}{a}\right)^b} \qquad (11)$$

The distribution is decided by several statistical parameter values, such as:

Negative of the log-likelihood:

$$NLogL = -\ln \prod_{i=1}^{N} f(x_i|\theta) \qquad (12)$$

where f is the theoretical probability density function with parameter θ (Bouyé et al., 2000).

Kolmogorov-Smirnov error:

$$KSE \ = \max\left|F_i - \hat{F}_i\right| \qquad (13)$$

Coefficient of determination:

$$R^2 = \frac{\Sigma_{i=1}^{n}(\hat{F}_i - \bar{F})^2}{\Sigma_{i=1}^{n}(\hat{F}_i - \bar{F})^2 + \Sigma_{i=1}^{n}(F_i - \hat{F}_i)^2} \qquad (14)$$

Chi-squared:

$$x^2 = \sum_{i=1}^{n} \frac{(F_i - \hat{F}_i)^2}{\hat{F}_i} \qquad (15)$$

Root mean squared error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(F_i - \hat{F}_i\right)^2} \qquad (16)$$

where $i = 1,2,\ldots,n$, $\hat{F}$ and F is the theoretical and empirical cumulative distribution function, respectively, and $\bar{F} = 1/n\, \Sigma_{i=1}^{n}\hat{F}_i$ (Kollu et al., 2012).

The third step is to identify the bias of training data. In this step, the relationship between HadISST and ECMWF data is calculated using a transfer function. The transfer function used is the linear transfer function because the linear function is better than the quadratic, cubic, or the difference of two quantiles for ECMWF sea surface temperature data (Nurdiati et al., 2019).

Let X and Y are the quantile values of the ECMWF and HadISST training data distribution, respectively, then the linear transfer function equation is

$$Y = f(X) = aX + b \qquad (17)$$

where a and b are constants (Najib & Nurdiati, 2021).

The next step, the ECMWF testing data, is identified to obtain the distribution function and CDF value. After that, the CDF data testing is corrected using the transfer function with the appropriate month and lead time to obtain the corrected prediction model CDF. The final step is to convert the corrected CDF into a corrected PDF prediction model using numerical derivatives.

Last step, the prediction models before and after correction were compared using their probability density functions, the mean and standard deviation ratio (std), and the expected error. Nurdiati et al. explained that the mean value of the probability density function (PDF) could be estimated using the Riemann sums approach (Nurdiati et al., 2019).

In a lecture note on the MAT 211 course of Arizona State University delivered by David Fishman (https://www.cengage.com/resource_uploads/downloads/1439049254_242719.pdf, accessed on March 14, 2021), suppose that the domain of probability density function f is a finite interval $[a, b]$. Break up the interval into n sub-intervals $[x_{i-1}, x_i]$, each of length Δx, as Riemann sums. Now, the probability of seeing a value of X in $[x_{i-1}, x_i]$ is approximate by $f(x_i)\Delta x$. Think of this as the fraction of times we expect to see values of X in this range. These values, all close to x_i, then contribute approximately $x_i f(x_i)\Delta x$ to the average, if we average together many observations of X. Adding together all of these contributions, we get

$$\text{mean}, \bar{x} = \textstyle\sum_{i=0}^{n} x_i (f(x_i) \Delta x_i^*) \quad (18)$$

Now, these approximations get better as $n \to \infty$, and the sum above is a Riemann sum converging to

$$E(X) = \int_a^b x f(x)\, dx \quad (19)$$

which is the formula of the mean or expected value of a continuous random variable X with probability density function f.

Using the same approach, the variance of X can be approximate using formula

$$\text{var}, \sigma^2 = \frac{\sum_{i=0}^{n} (x_i - \bar{x})^2 (f(x_i) \Delta x_i^*)}{\sum_{i=0}^{n} (f(x_i) \Delta x_i^*)} \quad (20)$$

the sum above is a Riemann sum converging to

$$\text{var}(X) = \int_a^b (x - \bar{x}) f(x)\, dx \quad (21)$$

which is the formula of variance value of a continuous random variable X. So, the standard deviation of can be approximate using formula

$$\text{std}, \sigma = \sqrt{\frac{\sum_{i=0}^{n} (x_i - \bar{x})^2 (f(x_i) \Delta x_i^*)}{\sum_{i=0}^{n} (f(x_i) \Delta x_i^*)}} \quad (22)$$

where x_i is the value of x-axis order-i, $f(x_i)$ is the probability density values of x_i and Δx_i is difference of x_i and x_{i-1}.

Based on that, the error of the model is given by

$$err = |\bar{x}_p - x_{obs}| \quad (23)$$

with $\bar{x}_p$ is the mean value of the prediction model and x_{obs} is the actual Nino index. After that, the value of mean and standard deviation ratio is given by

$$\text{mean ratio}, r\bar{x} = \frac{(\bar{x}_{cor} - x_{obs})}{|\bar{x}_{mod} - x_{obs}|} \quad (24)$$

$$\text{std ratio}, r\sigma = \frac{\sigma_{cor}}{\sigma_{mod}} \quad (25)$$

with $\bar{x}_{cor}$ is the mean value of the corrected ECMWF model, $\bar{x}_{mod}$ is the mean value of the ECMWF model and x_{obs} is the actual value, and σ_{cor} is the standard deviation value of the corrected ECMWF model and σ_{mod} is the standard deviation value of the ECMWF model. Moreover, the expected error value is the average error that occurs for each ECMWF data correction in each month and lead time.

RESULTS AND DISCUSSION

Data Extraction

The Hadley Centre Global Sea Ice and Sea Surface Temperature (HadISST) is monthly global sea surface temperature data. HadISST data is downloaded in the Network Common Data Form (NetCDF), composed of four variables: lon, *lat*, *time,* and *sst*. The time vector variable has a size of 1798 × 1, the latitude (*lat*) vector has a size of 180 × 1, the longitude (*lon*) vector has a size of 360 × 1, and the *sst* variable shows sea surface temperature data which has a size of 360 × 180 × 1798 (longitude × latitude × time) in Celcius degree (°C). The data extraction process begins with reading the netCDF data file. Then the data is cut from 1983-2012 for the HadISST training data. Lastly, the data is truncated for the Nino 3.4 region, which is 5°S-5°N, and 170°W-120°W (K. E. Trenberth, 1997).

The ECMWF prediction data used is sea surface temperature prediction data with ensemble modeling. Prediction data is downloaded around Nino 3.4, which has dimensions with five variables: lon, *lat*, *ensemble*, *lead time,* and *sst*. The lon and lat variables show the longitude and latitude of the ECMWF prediction data, which has a size of 205 × 1 and 45 × 1, respectively. The ensemble variable indicates the number of predictive models from the ECMWF prediction data and has a size of 25 × 1. The lead time variable shows the prediction period, which is 216 days or approximately seven months. The sst variable shows the sea surface temperature from the ECMWF prediction data, which has a size of 205 × 41 × 25 × 216 (longitude × latitude × ensemble × lead time). The ECMWF prediction data has a spatial resolution of 0.25°× 0.25°. The ECMWF prediction data that will be used is the ECMWF forecast data for the years 1982-2018 in Nino 3.4 area in the NetCDF (Network Common Data Form) format. The ECMWF data extraction process begins with reading the netCDF data file. Each data file has its own "add_offset" and "scale_factor" values and is in Kelvin units. Therefore, the data file is transformed into units of degrees Celsius (°C) using the following equation

$$T^* = T \times \text{SF} + \text{AO} - 273.15 \quad (26)$$

SF and AO are the "scale factor" dan "add offset" attribute from each NetCDF file, respectively, is the initial sea surface temperature and the sea surface temperature that is ready to be processed and used. After that, the data was readjusted for its location in the Nino 3.4 area and the data was compiled into training data from 1983-2012 and testing data from 2013-2018.

Distribution Identification

Distribution identification is carried out on the extracted HadISST and ECMWF training data. Distribution identification is done every month and lead times so that there are 12 distributions for 12 months on HadISST data and 12 × 7 distributions for 12 months and 7 lead times on ECMWF training data. The data distribution is identified using several distributions that have been mentioned in the research step. After several distribution functions were obtained, the distribution functions were tested using the goodness-of-fit test with several parameters, which are KSE, NLogL, R^2, and x^2. Kollu et al., (2012) said the goodness-of-fit test was used to measure the deviation between the predicted data using the theoretical distribution function and the empirical data distribution. The smaller KSE, NLogL, x^2, and RMSE values and the larger R^2 value resulted in the distribution getting closer to the actual data. The results of the process of identifying the distribution of HadISST and ECMWF training data can be seen in Figure 1.

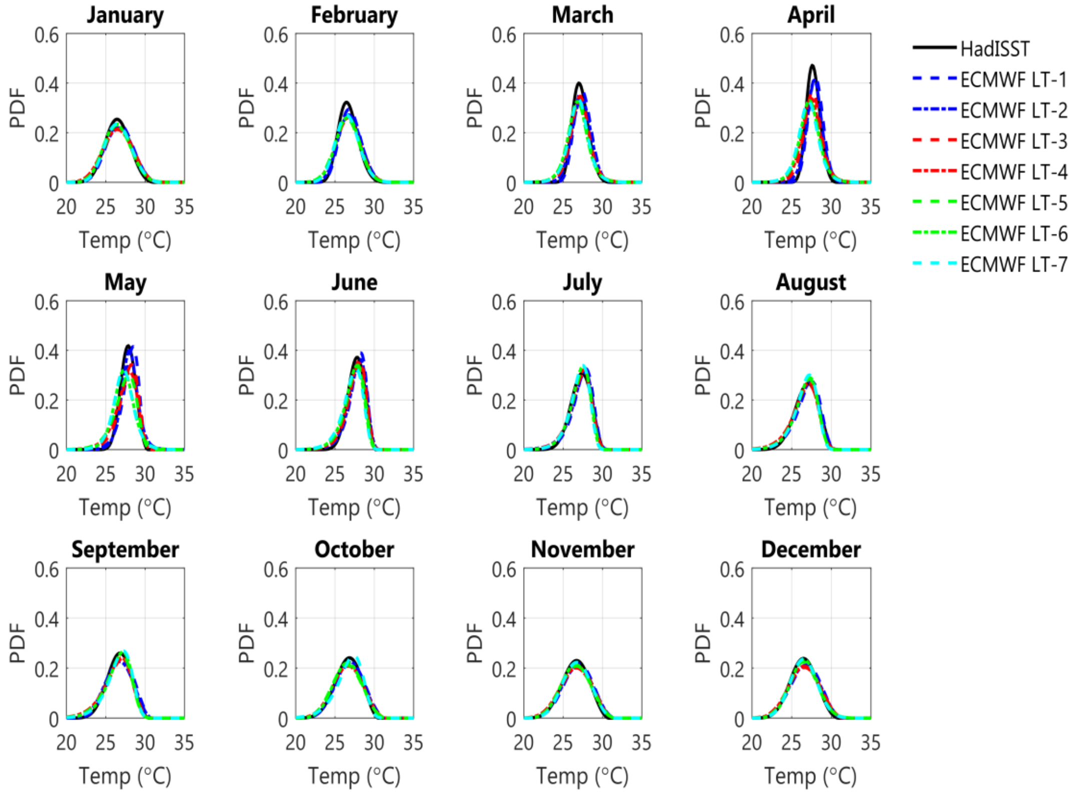


Figure 1. The results of distribution identification of HadISST and ECMWF training data for each month and lead time

Bias Identification

After the distribution of HadISST and ECMWF training data are obtained, bias identification was carried out to find a transfer function between the HadISST and ECMWF training data. Bias is the difference or deviation between the expected value of the estimator and the predicted parameter. Bias identification can be observed by looking at the HadISST, and ECMWF training data distribution function's different quantile values.. A quantile is a value that divides a range of data into equal parts. The n-th quantile value of X is given by

$$Q(n) = F^{-1}(n) \text{ for } 0 < n < 1 \quad (27)$$

where $Q(n)$ is the n-th quantile value of X and $F(n)$ is the n-th CDF value of X. The quantity to be used is 40 parts with an upper and a lower limit of 0.001 and 0.999. Thus, the value of n that will be used is 0.001, 0.025, 0.05, ..., 0.975, and 0.999.

Determining the transfer function starts with calculating the quantile value of the HadISST and ECMWF training data distribution. After that, calculate the linear regression function between the ECMWF data quantiles against the HadISST data. After the linear equations are obtained, calculate the R^2 and RMSE values to test the goodness-of-fit of the transfer function against the data. The results of the transfer function in the first lead time can be seen in Figure 2.

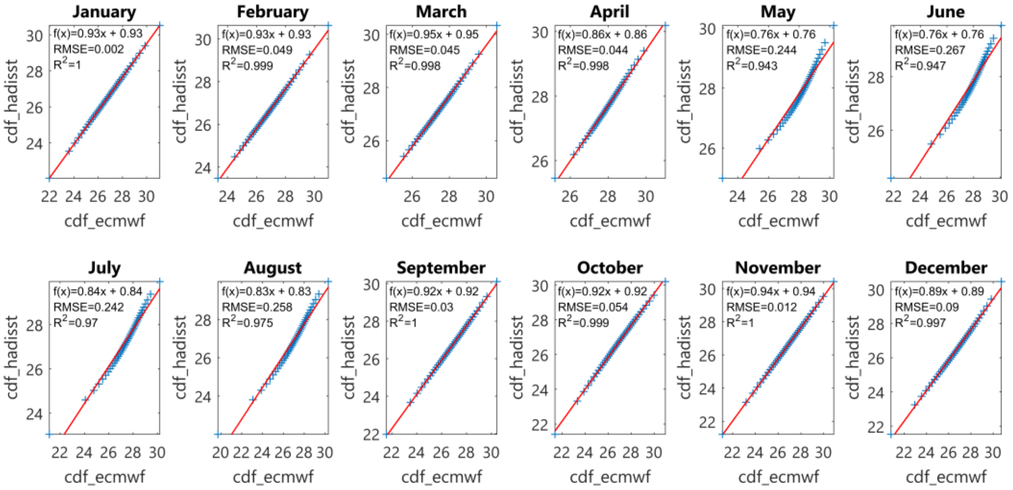


Figure 2. Plot of the results of determining the transfer function between the HadISST and ECMWF training data at the 1st lead time from January to December using the linear transfer function.

The transfer function can be said to be good and acceptable if it has an R^2 value more than 70% (Dasanto et al., 2014). Based on Figure 2, it can be seen that the linear function is acceptable as a transfer function for sea surface temperature data in the Nino 3.4 area because the R^2 value of the transfer function is more than 90%. Because the transfer function is acceptable, the transfer function can be used for further processing. The process of determining the transfer function is repeated for the 2nd to 7th lead times and the value of the goodness of fit test shows that the transfer function is acceptable.

Bias Correction for Testing Data

In this step, bias correction is carried out from 2013-2018 ECMWF testing data. The bias correction step begins with data extraction. The next step is to identify the distribution of testing data for each month, lead time, and year. After the distribution identification results are obtained for each month, lead time, and year, the step is continued by calculating the CDF and quantile values with predetermined n values. Next, correct the quantile of the ECMWF data by evaluating the value on the transfer function with the appropriate month and lead times, so that the CDF of the corrected ECMWF data is obtained. The results of the bias correction towards the ECMWF testing data in 2013 at the 1st lead time can be seen in Figure 3.

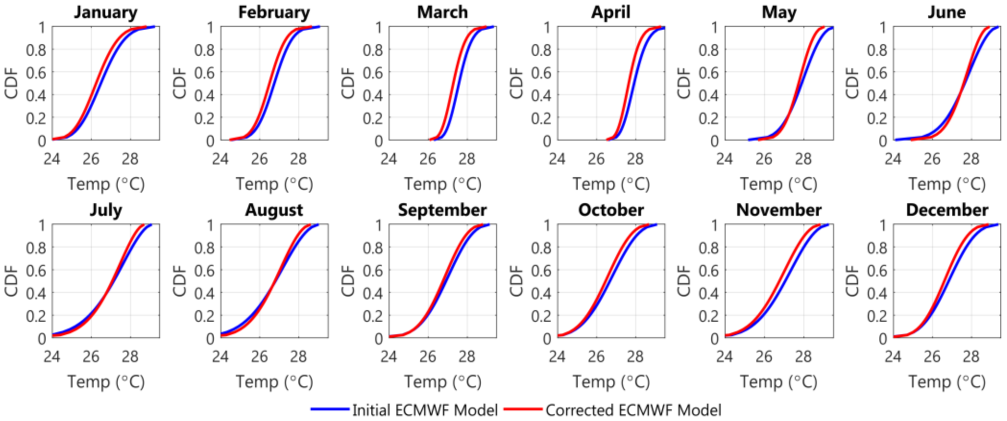


Figure 3. The CDF results of the bias correction step towards the ECMWF testing data in 2013 at the 1st lead time

Figure 3 shows that there was a shift in the CDF to the left, meaning that the 2013 ECMWF data at the first lead time were generally overestimated. The correction process is continued for the 1st to 7th lead time from 2013-2018. After the CDF of the corrected ECMWF data is obtained, the PDF of the corrected ECMWF data is calculated by deriving the CDF using a numerical derivative, which is the finite difference method. The PDF results of the bias correction step towards the ECMWF testing data in 2013 at the 1st lead time and their comparison to the exact values can be seen in Figure 4.

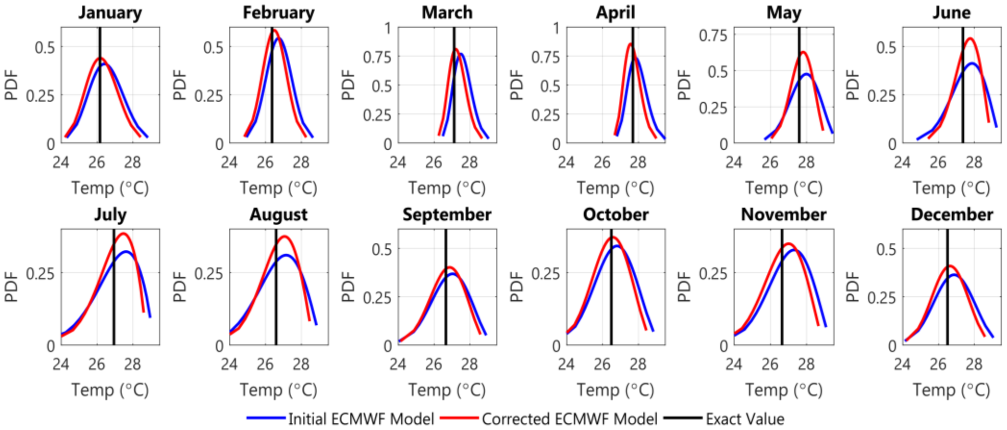


Figure 4. The PDF results of the bias correction step towards the ECMWF testing data in 2013 at the 1st lead time and their comparison to the exact values

Figure 4 shows that the PDF of the corrected ECMWF model higher than the initial ECMWF model, which means there has been an improvement in the standard deviation of the ECMWF prediction data. Also, it can be seen that there is a shift in the graph towards the left of the initial ECMWF as in Figure 3. The PDF calculation process is repeated for each month and lead times from 2013-2018.

Evaluation of The Prediction Models

Comparison of the PDF

This section will compare the PDF value of the ECMWF data before and after correction against the exact El Nino index value. Comparison of the PDF value of ECMWF data before and after correction to the exact value of the El Nino index at the 1st lead time in 2013 can be seen in Figure 4. Based on Figure 4, almost all months can be corrected for the accuracy and standard deviation values shown by the PDF function of the corrected model which is closer to the Nino index value and higher than the PDF function of the initial ECMWF model. For more details, the comparison of the mean value of the PDF function at the 1st lead time in 2013 can be seen in Figure 5.

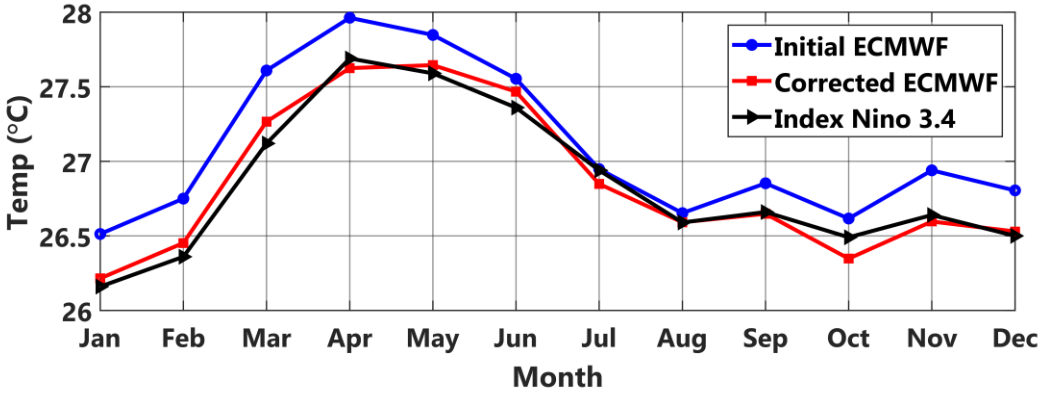


Figure 5. The comparison of the mean value of the PDF function of the ECMWF model before and after correction to the exact Nino index value at the 1st lead time in 2013

Based on Figure 5, it can be seen that the mean of the corrected ECMWF model is closer to the exact value of the Nino 3.4 index than the mean of the initial ECMWF model, meaning that the corrected model has better accuracy than the initial model.

The mean and standard deviation ratio

This section will compare the accuracy and precision of the ECMWF model before and after correction based on the mean ratio ($r\bar{x}$) and the standard deviation ratio ($r\sigma$). The mean and standard deviation ratio values are given by equations (24) and (25). The mean ratio is used to compare the accuracy of the ECMWF model before and after correction, while the standard deviation ratio is used to compare the standard deviation of the ECMWF model before and after correction.

Two conditions for the corrected model can be said to be better than the initial model, which are the mean ratio value must be in the interval $(-1,1)$ and the standard deviation ratio value is in the interval $(0,1)$ (Lealdi et al., 2018; Nurdiati et al., 2019). For example, the calculation of the mean and standard deviation ratio at the first lead time in 2013 can be seen in Table 1.

Table 1. The value of the mean and standard deviation ratio at the 1st lead time in 2013

Month	$\bar{x}_{mod}$	$\bar{x}_{cor}$	x_{obs}	$r\bar{x}$	σ_{mod}	σ_{cor}	$r\sigma$
January	26.51	26.21	26.16	0.1548	0.9471	0.8933	0.9431
February	26.75	26.45	26.36	0.2351	0.7356	0.6967	0.9471
March	27.61	27.27	27.12	0.2982	0.5220	0.5046	0.9667
April	27.96	27.63	27.69	-0.2373	0.5496	0.4820	0.8770
May	27.85	27.65	27.59	0.2161	0.8038	0.6184	0.7694
June	27.55	27.47	27.36	0.5537	0.9504	0.7350	0.7734
July	26.95	26.85	26.94	-11.571	1.2897	1.0990	0.8520
August	26.65	26.59	26.59	0.0141	1.2807	1.0755	0.8398
September	26.85	26.65	26.66	-0.0614	1.0387	0.9635	0.9276
October	26.62	26.35	26.49	-1.1272	1.1197	1.0411	0.9298
November	26.94	26.60	26.64	-0.1444	1.1904	1.1318	0.9508
December	26.80	26.53	26.50	0.0960	1.0494	0.9437	0.8993

Table 1 shows that the results of statistical bias correction using quantile mapping approach at the 1st lead time in 2013 all months except July and October get satisfactory results which is the accuracy and standard deviation value of the initial model can be corrected indicated by the mean ratio values are in the interval $(-1,1)$ and the standard deviation ratio values are in the interval $(0,1)$. Meanwhile, July and October can be said to be quite good because only the standard deviation value can be corrected. The results of the bias correction evaluation can be visualized in a scatter plot of the ratio of the mean to standard deviation of the ECMWF data ratio each month and lead time from 2013-2018 which can be seen in Figure 6. Each point in Figure 6 represents the mean and standard deviation ratio for a given month, lead time, and year.

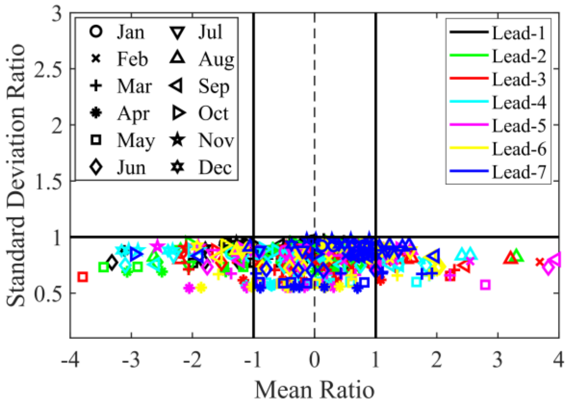
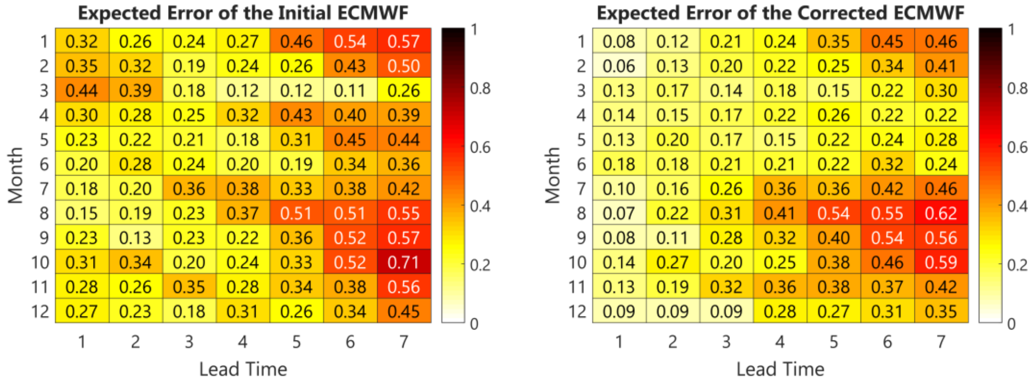


Figure 6. The plot of the mean ratio against standard deviation ratio of the ECMWF data for each month and lead time in 2013-2018

Based on Figure 6, the standard deviation value of all ECMWF testing data in 2013-2018 can be corrected using the quantile mapping approach, which is indicated by all points in Figure 6 as less than 1 on the y-axis. Meanwhile, the accuracy value of most of the ECMWF testing data in 2013-2018 (about 60% of the data) can be corrected using the quantile mapping method, which is shown by the points in Figure 6 being in the interval (-1, 1) on the x-axis.

Expected error

The expected error is the expected value of the error between the true value and the resulting value. The expected error value can be found by calculating the average accuracy value of the ECMWF data before and after correction in 2013-2018 for each month and lead time. The comparison of the expected error value from the ECMWF data before and after correction for each month and the lead time can be seen in Figure 7. Sebagai contoh, dapat dilihat pada Gambar 1.



Expected Error of the Initial ECMWF

Month	1	2	3	4	5	6	7
1	0.32	0.26	0.24	0.27	0.46	0.54	0.57
2	0.35	0.32	0.19	0.24	0.26	0.43	0.50
3	0.44	0.39	0.18	0.12	0.12	0.11	0.26
4	0.30	0.28	0.25	0.32	0.43	0.40	0.39
5	0.23	0.22	0.21	0.18	0.31	0.45	0.44
6	0.20	0.28	0.24	0.20	0.19	0.34	0.36
7	0.18	0.20	0.36	0.38	0.33	0.38	0.42
8	0.15	0.19	0.23	0.37	0.51	0.51	0.55
9	0.23	0.13	0.23	0.22	0.36	0.52	0.57
10	0.31	0.34	0.20	0.24	0.33	0.52	0.71
11	0.28	0.26	0.35	0.28	0.34	0.38	0.56
12	0.27	0.23	0.18	0.31	0.26	0.34	0.45

Expected Error of the Corrected ECMWF

Month	1	2	3	4	5	6	7
1	0.08	0.12	0.21	0.24	0.35	0.45	0.46
2	0.06	0.13	0.20	0.22	0.25	0.34	0.41
3	0.13	0.17	0.14	0.18	0.15	0.22	0.30
4	0.14	0.15	0.17	0.22	0.26	0.22	0.22
5	0.13	0.20	0.17	0.15	0.22	0.24	0.28
6	0.18	0.18	0.21	0.21	0.22	0.32	0.24
7	0.10	0.16	0.26	0.36	0.36	0.42	0.46
8	0.07	0.22	0.31	0.41	0.54	0.55	0.62
9	0.08	0.11	0.28	0.32	0.40	0.54	0.56
10	0.14	0.27	0.20	0.25	0.38	0.46	0.59
11	0.13	0.19	0.32	0.36	0.38	0.37	0.42
12	0.09	0.09	0.09	0.28	0.27	0.31	0.35

Figure 7. The comparison of the expected error value from the ECMWF data before and after correction for each month and the lead time

Based on Figure 7, the expected error of the corrected ECMWF is brighter than the expected error of the initial ECMWF. This shows that the quantile mapping method can correct the expected error of the ECMWF prediction data around the Nino 3.4 area almost every month and lead time, especially at the 1st to 4th lead time in all months and at the 5th to 7th lead time in November to June. Prediction for July-October is difficult to correct because sea surface temperature in the Nino 3.4 area is very fluctuating during these months. In climatology, the difficulty of climate prediction from July to October is known as the "spring predictability barrier".

CONCLUSIONS AND SUGGESTIONS

Using the quantile mapping approach, the statistical bias correction method can adequately correct the ECMWF testing model in most of the lead times and months based on the HadISST and ECMWF training data. For this case, the results showed that 60% of El Nino's predictions on the testing data had improved the mean value. Also, all of El Nino's predictions on the testing data have improved the standard deviation value. Moreover, the quantile mapping approach is effectively used to correct ECMWF prediction data's bias for all months in the 1st to 4th lead times. But, in the 5th to 7th lead times, only November-June can be corrected. This is indicated by the expected error of the corrected ECMWF model is smaller than the expected error of the initial ECMWF model.

This study is only limited to using a quantile mapping approach with a linear transfer function. Many transfer functions other than linear can be used, such as polynomials, exponential, and even only the difference of each quantile point (not an equation). Besides the quantile mapping approach, there are other approaches for statistical bias correction, such as mean and variance scaling, nudging bias correction, change factor bias correction, etc. Various types of transfer functions and other approaches can be used to correct ECMWF data bias to predict the El Nino index and can be compared with the quantile mapping approach in this study.

REFERENCES

Aldrian, E., & Susanto, R. D. (2003). Identification of three dominant rainfall regions within indonesia and their relationship to sea surface temperature. *International Journal of Climatology*, *23*(12), 1435–1452. https://doi.org/10.1002/joc.950

Ayugi, B., Tan, G., Ruoyun, N., Babaousmail, H., Ojara, M., Wido, H., Mumo, L., Ngoma, N. H., Nooni, I. K., & Ongoma, V. (2020). Quantile mapping bias correction on rossby centre regional climate models for precipitation analysis over Kenya, East Africa. *Water (Switzerland)*, *12*(3), 801. https://doi.org/10.3390/w12030801

Bouyé, E., Durrleman, V., Nikeghbali, A., Riboulet, G., & Roncalli, T. (2000). Copulas for finance - a reading guide and some applications. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1032533

Chhikara, R. S., & Folks, J. L. (1977). The inverse gaussian distribution as a lifetime model. *Technometrics*, *19*(4), 461–468. https://doi.org/10.1080/00401706.1977.10489586

Dasanto, B. D., Boer, R., Pramudya, B., & Suharnoto, Y. (2014). Evaluasi curah hujan TRMM menggunakan pendekatan koreksi bias statistik. *Jurnal Tanah Dan Iklim*, *38*(1), 15–24. https://doi.org/10.2017/jti.v38i1.6244

Decani, J. S., & Stine, R. A. (1986). A note on deriving the information matrix for a logistic distribution. *American Statistician*, *40*(3), 220–222. https://doi.org/10.1080/00031305.1986.10475398

E.J. Gumbel. (1941). Return Period of Floods. *The Annals of Mathematical Statistics*, *12*(2), 163–190.

Hald, A. (1949). Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point. *Scandinavian Actuarial Journal*, *1949*(1), 119–134. https://doi.org/10.1080/03461238.1949.10419767

Hannachi, A. (2004). A primer for EOF analysis of climate data. *Department of Meteorology, University of Reading*, *1*(29).

Hosking, J. R. M., Wallis, J. R., & Wood, E. F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, *27*(3), 251–261.

Katiraie-Boroujerdy, P. S., Naeini, M. R., Asanjan, A. A., Chavoshian, A., Hsu, K. lin, & Sorooshian, S. (2020). Bias correction of satellite-based precipitation estimations using quantile mapping approach in different climate regions of Iran. *Remote Sensing*, *12*(13), 2102. https://doi.org/10.3390/rs12132102

Kollu, R., Rayapudi, S. R., Narasimham, S. V. L., & Pakkurthi, K. M. (2012). Mixture probability distribution functions to model wind speed distributions. *International Journal of Energy and Environmental Engineering*, *3*(1), 1–10. https://doi.org/10.1186/2251-6832-3-27

Lealdi, D., Nurdiati, S., & Sopaheluwakan, A. (2018). Statistical bias correction modelling for seasonal rainfall forecast for the case of Bali island. *Journal of Physics: Conference Series*, *1008*(1), 12–18. https://doi.org/10.1088/1742-6596/1008/1/012018

Misnawati, Boer, R., June, T., & Faqih, A. (2018). Perbandingan metodologi koreksi bias data curah hujan CHIRPS. *LIMNOTEK - Perairan Darat Tropis Di Indonesia*, *25*(1), 18–29. https://limnotek.limnologi.lipi.go.id/index.php/limnotek/article/view/224

Nabilah, F., Prasetyo, Y., & Sukmono, A. (2017). Analisis pengaruh fenomena el nino dan la nina terhadap curah hujan tahun 1998 - 2016 menggunakan indikator oni (Oceanic Nino Index) (Studi Kasus : Provinsi Jawa Barat). *Jurnal Geodesi Undip*, *6*(4), 402–412.

Najib, M. K., & Nurdiati, S. (2021). Koreksi bias statistik pada data prediksi suhu permukaan air laut di wilayah indian ocean dipole barat dan timur. *Jambura Geoscience Review*, *3*(1), 9–17. https://doi.org/10.34312/jgeosrev.v3i1.8259

Nurdiati, S., Sopaheluwakan, A., & Najib, M. K. (2019). Statistical bias correction for predictions of indian ocean dipole index with quantile mapping approch. *International MIPAnet Conference on Science and Mathematics (IMC-SciMath), Medan*.

Passow, C., & Donner, R. V. (2020). Regression-based distribution mapping for bias correction of climate model outputs using linear quantile regression. *Stochastic Environmental Research and Risk Assessment*, *34*(1), 87–102. https://doi.org/10.1007/s00477-019-01750-7

Pastén-Zapata, E., Jones, J. M., Moggridge, H., & Widmann, M. (2020). Evaluation of the performance of Euro-CORDEX Regional Climate Models for assessing hydrological climate change impacts in Great Britain: A comparison of different spatial resolutions and quantile mapping bias correction methods. *Journal of Hydrology*, *584*, 124653.

Philander, S. G. (1989). El Niño, La Niña, and the Southern Oscillation. *International Geophysics Series*, *46*, X–289.

Piani, C., Haerter, J. O., & Coppola, E. (2010). Statistical bias correction for daily precipitation in regional climate models over Europe. *Theoretical and Applied Climatology*, *99*(1), 187–192. https://doi.org/10.1007/s00704-009-0134-9

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., Kent, E. C., & Kaplan, A. (2003). Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research: Atmospheres*, *108*(D14), 4407. https://doi.org/10.1029/2002jd002670

Shonk, J. K., Demissie, T. D., & Toniazzo, T. (2019). A double ITCZ phenomenology of wind errors in the equatorial Atlantic in seasonal forecasts with ECMWF models. *Atmospheric Chemistry and Physics*, *19*(17), 11383–11399.

Taufik, & Marnita. (2004). *IPBA (Imu pengetahuan bumi dan antariksa)*. Universitas Al Muslim.

Trenberth, K. E. (1997). The definition of el niño. *Bulletin of the American Meteorological Society*, *78*(12), 2771–2777. https://doi.org/10.1175/1520-0477(1997)078<2771:TDOENO>2.0.CO;2

Trenberth, K., & National Center for Atmospheric Research Staff. (2020). *The climate data guide: Nino SST indices (Nino 1+2, 3, 3.4, 4; ONI and*

TNI). https://climatedataguide.ucar.edu/climate-data/nino-sst-indices-nino-12-3-34-4-oni-and-tni

Walpole, R. E. (1990). P*engantar statistika, edisi ke-3 (Introduction to statistics)*. PT Gramedia Pustaka Utama.