Machine learning improves the modelled wave spectrum in the North Sea

M.P. de Ridder[1], J.P. den Bieman[2], C. Gautier[3] and J.C.C. van Nieuwkoop[4]
[1]Department of Coastal structures and waves, Deltares, Boussinesqweg 1, 2629 HV, Delft, Netherlands. Menno.deridder@deltares.nl.
[2]Department of Coastal structures and waves, Deltares, Boussinesqweg 1, 2629 HV, Delft, Netherlands. Joost.denbieman@deltares.nl.
[3]Department of Coastal structures and waves, Deltares, Boussinesqweg 1, 2629 HV, Delft, Netherlands. Caroline.Gautier@deltares.nl.
[4]Department of Coastal structures and waves, Deltares, Boussinesqweg 1, 2629 HV, Delft, Netherlands. Joana.Nieuwkoop@deltares.nl.

Abstract

To improve the energy density and directional spectra computed with the SWAN model for the North Sea, a data-driven model is trained to correct the SWAN spectra. After training of the data-driven model on a year of observed and modelled data, the energy density and directional spectrum are corrected for three locations in the North Sea. When this correction is applied, the SWAN results are significantly improved. Both the energy density and the directions show a reduction in RMSE of up to 30% for the directions and 26 % for the energy density. Due to the short computational time of the data-driven model, this approach can easily be implemented in an operational forecast system.

1. Introduction

Nearshore wave predictions are important for safety measures against flooding, ship navigation and the design of coastal flood defences (Thomas and Dwarakish, 2015). Numerical wave models are applied to transform the wave properties from deep water to coastal shallow water. Due to model simplifications and errors in model input, e.g. bathymetry and boundary conditions, the model predictions are generally not perfect (Gautier and Caires, 2015; Van Dongeren et al., 2011). Traditionally, to improve these wave predictions, either a more sophisticated model could be applied or the model input can be improved (e.g. higher resolution of wind field). However, this would mean that the computational time would increase or more data is required. An alternative is to estimate the error with a data-driven model given the hydrodynamic conditions and correct for that error. Beforehand, this method requires training on a large dataset with observations to learn the error patterns of the numerical model as a function of the hydrodynamic conditions and the location. After training, the data-driven model can be used during operational forecasts with negligible computational effort.

In Callens et al. (2020) this approach is applied to correct the wave parameters computed with the SWAN model for locations near the French coast. Based on the correction with the data-driven model, they were able to reduce the RMSE error by 8 to 10%. In this study, we extended this approach to the correction of the energy density spectrum and directional spectrum obtained from the spectral wave model SWAN (Booij et al., 1997) for the North Sea schematization (Deltares, 2018). Based on one year of operational SWAN results and wave observations, the data-driven model has been trained to correct the wave spectra predicted by SWAN.

2. Method

2.1 *Study site and data*

The 1D energy density and directional spectra from the SWAN North Sea schematization are corrected with the data-driven model. The model domain includes the Dutch North Sea and the Waddensea. Since this model runs operationally, the model results are available for a long period of time. In this study three characteristic locations for shipping navigation (Europlatform, Eurogeul E13 and stroompaal IJmond) are selected to show the performance of the correction with a data-driven model (See Figure 1). Besides the model forecasts, the observed wave parameters, wind parameters and wave spectra are collected for the period January 1st, 2019 to August 1st, 2020 with an hourly temporal resolution. This dataset contains various wave and wind conditions with a wave height variation ranging from 0.1 to 5.0 m and wind speeds of up to 21 m/s. In this way the data-driven model is capable of predicting the correction for different wave conditions. The modelled and observed wave spectra are interpolated to a frequency axis with 28 frequencies to define a correction for each frequency bin. Moreover, data entries with missing or unrealistic values were removed from the dataset. In total the dataset for the three locations and 28 frequencies for the energy density correction consists of 480,480 entries and the dataset for the directional dataset consist of 482,888 entries.

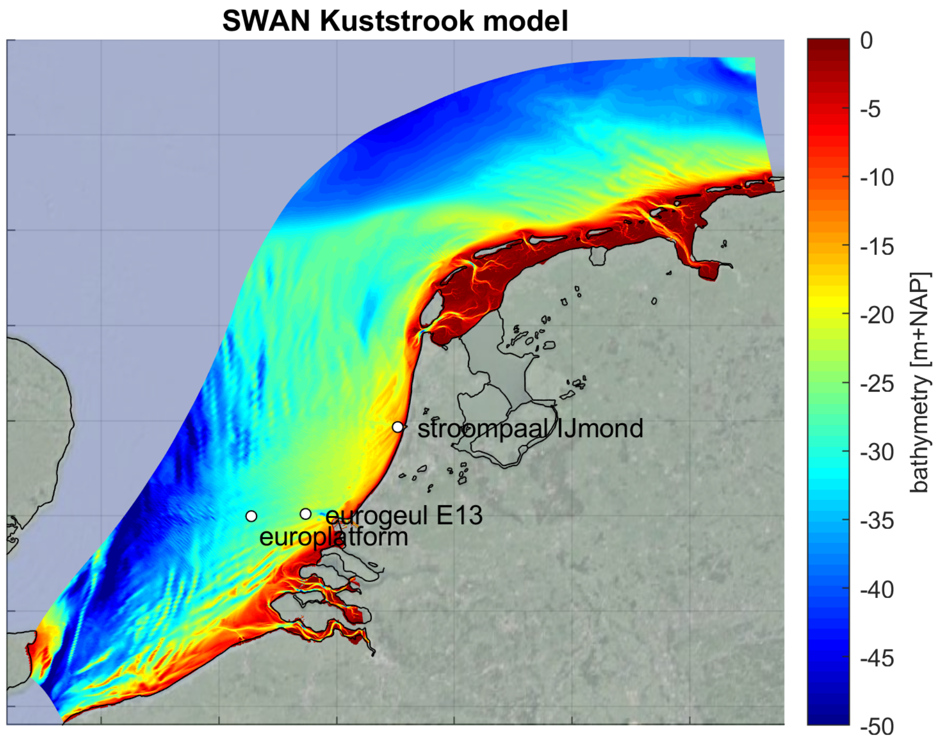

Figure 1. SWAN North Sea schematization with the three locations.

2.2 *General framework*

The data-driven model is applied to correct the SWAN results, which means that both models are required to obtain the corrected forecast of the wave spectrum. The general approach is to first compute the wave parameters and wave spectra with SWAN. Based on the wave parameters obtained from SWAN and the wind parameters used as SWAN input, the data-driven model computes the correction for each location and frequency bin. Both a correction of the energy density and direction for each frequency bin is computed. When combing the correction and the initial wave spectrum, a reconstructed improved wave spectrum is obtained (see Figure 2). The data-driven model needs to be trained once on a large dataset, which is time consuming. However, the prediction of the correction is very fast.

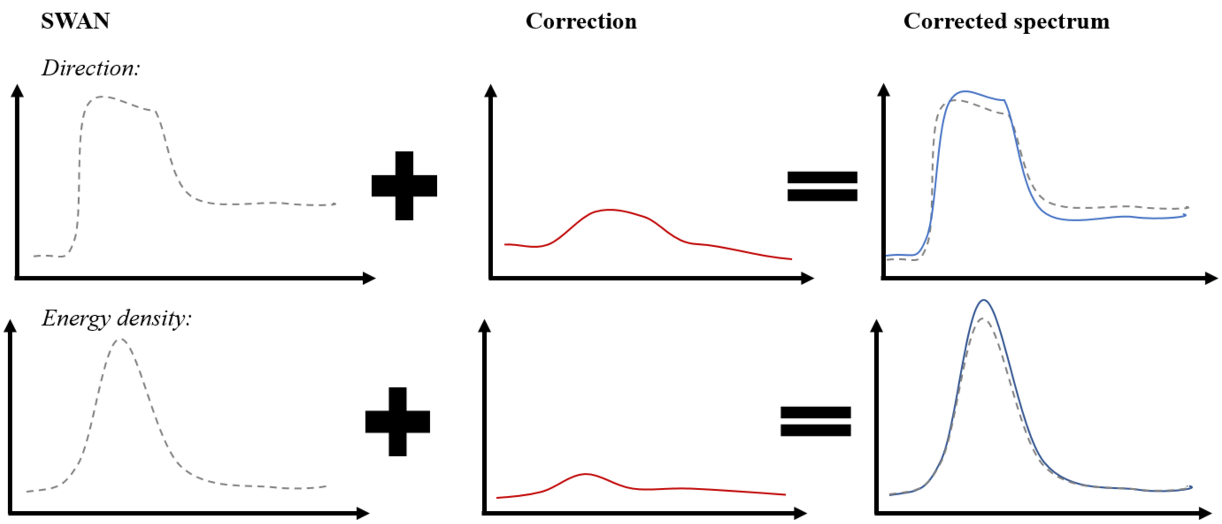

Figure 2. Overview of the approach for the correction of the energy density spectrum and directional spectrum.

2.3 *Data-driven model*

The gradient boosting method, XGBoost (Chen and Guestrin, 2016), is applied as data-driven model to predict the correction. This supervised learning method optimizes a large set of decision trees, which predicts a target output variable given a set of input features. The training settings of the model are shown in Table 1. To prevent overfitting the training dataset is divided into a test and training dataset. When the error of the test set is not decreasing every *early stopping rounds* the training is finished. No further optimization of these hyperparameters is applied.

Table 1. Applied hyperparameter settings in the XGBoost model.

Parameter	**Value**
Learning rate	0.0075
Max depth	15
Minimum child weight	5
lambda	1
subsample	1
Early stopping rounds	1000

Two different data-driven models were set up. These two data-driven models give the error correction as output given a set of input variables. The first data-driven model is applied to correct the energy density in a given frequency bin with the normalized energy density in a frequency bin as error metric,

$$\Delta E(f) = \frac{E_{obs}(f) - E_{SWAN}(f)}{E_{SWAN,total}} \tag{1.1}$$

Where $E_{SWAN}(f)$ is the computed energy density by SWAN, $E_{obs}(f)$ the observed energy density and $E_{SWAN,total}$ the total amount of energy in the energy density spectrum. The second data-driven model is applied to correct the direction in a given frequency bin with the normalized directional error:

$$\Delta D(f) = (D_{OBS}(f) - D_{SWAN}(f)) / 360 \tag{1.2}$$

Where $D_{SWAN}(f)$ is the computed direction by SWAN and $D_{obs}(f)$ the observed direction. The directional error is limited between -180° and +180° North. The physical input variables for both models are the SWAN computed wave height, spectral period, wave direction, wind speed, wind direction and water level. In this way, no observations are required during operational forecasts since the data-driven model only depends on numerical results. Observations are only required during the training stage of the data-driven model to quantify the error between forecast and observation. Next to these physical input parameters, the models have a location index and frequency bin index as input variables.

The dataset is randomly divided into a training set (75%) and a validation set (25%). Due to the random division of the dataset, the training set is representative of the validation set. The training set is divided into a pure training set (75%) and a testing set required for the stop criteria (25%).

3. Results

The trained data-driven model is applied to predict the correction of the energy density and the directions for the frequency bins. Note that the data from the validation period has not been seen by the model before and, therefore, it is a good dataset to check the performance. When the SWAN predictions are corrected with the data-driven model, the errors with respect to the observations are significantly reduced (See Table 1). Circular statistics are applied to compute the RMSE for the directions. For the three locations, a reduction of 20 to 30% is found for the RMSE in the energy density spectrum and a reduction of 20 to 36% is found for the direction.

In Figure 3 and Figure 4 the scatter plots for Europlatform are shown. Both the scatter cloud for the energy density and direction contains less scatter after applying the correction. The scatter plot for the directions obtained with SWAN shows bands around 150 and 220°, which is caused by the directions from the lower frequency bins in SWAN. These frequency bins in SWAN appear to have a preferred direction. After applying the corrections, these bands disappear and a better prediction for the direction is found. Note that the scatter plots for the directions also show points in the upper left and lower right corner since these directions corresponds to respectively 0 and 360° degree. Similar results were found for the other two locations.

Table 1. Root mean squared errors (RMSE) of the SWAN results and the corrected SWAN results for both the energy density and direction.

Location	RMSE energy density [m^2/Hz]		RMSE direction [°]	
	SWAN	SWAN+correction XGB	SWAN	SWAN+correction XGB
E13	0.42	0.33	20.57	6.26
EPL	0.49	0.34	19.16	14.81
Stroompaal IJmond	0.48	0.45	20.06	12.87

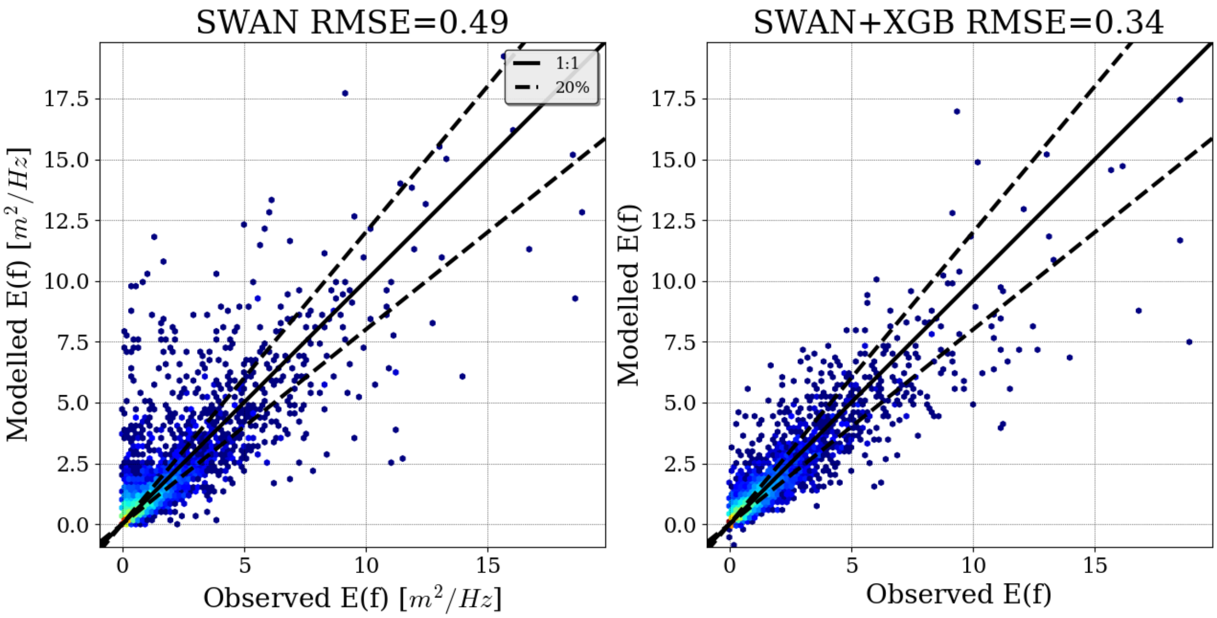

Figure 3. Scatter plot of the energy density between the observation and the SWAN results (left panel) and the observations and the corrected SWAN results (right panel). The colors indicate the density of the points. The solid black line represents perfect agreement and the dashed lines show the 20% deviations. The results for location Europlatform are shown.

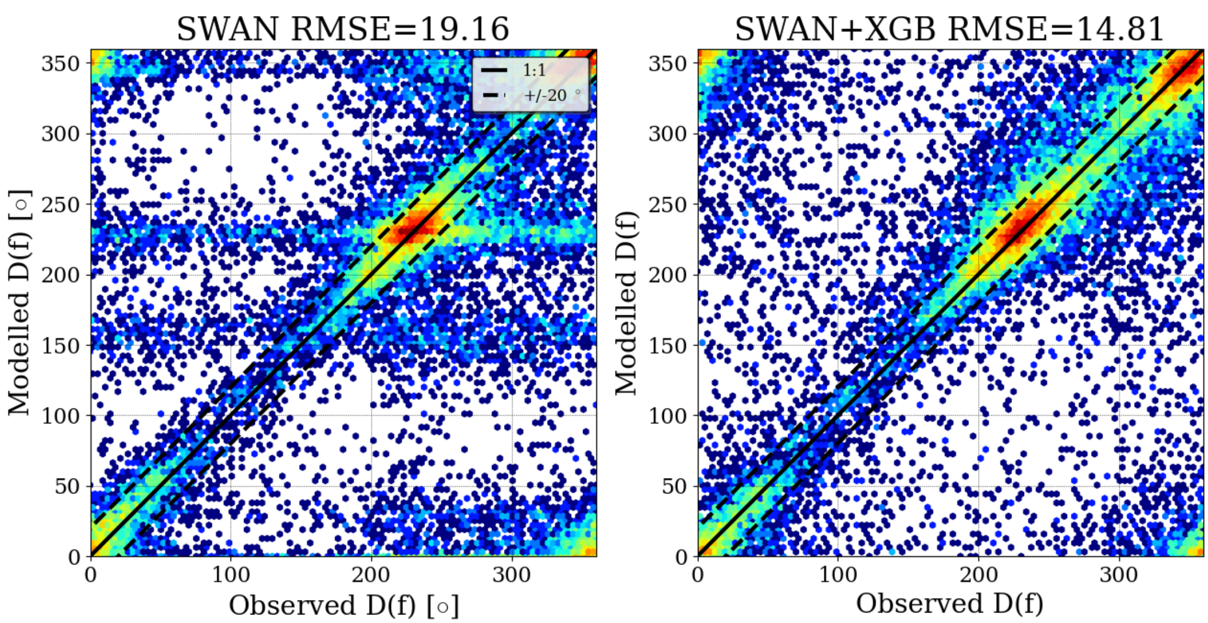

Figure 4. Scatter plot of the directions between the observation and the SWAN results (left panel) and the observations and the corrected SWAN results (right panel). The colors indicate the density of the points. The solid black line represents perfect agreement and the dashed lines show the absolute 20° deviation. The results for location Europlatform are shown.

Apart from the energy density spectrum and directional spectrum, the wave parameters based on the reconstructed energy density spectrum are computed. Since the correction is derived per single frequency bin, it is not guaranteed that a realistic spectral shape is obtained after the correction. However, the RMSE of the wave parameters is even reduced when the corrected spectrum is compared to the original spectrum. The improvement of the error in wave height shows that the total energy within the spectrum is better captured after the correction. The spectral period and mean absolute wave period indicate that the spectral shape is also better captured after the correction, where the mean absolute wave period T_{m02} is sensitive for the high frequency tail of the spectrum and the spectral period ($T_{m-11,0}$) is more representative for the energy at the lower frequencies. To show the performance for the low-frequency waves, important for shipping navigation, the low-frequency wave height based on the total energy below 0.1 Hz is computed. Similar to the other wave parameters, the reconstructed spectrum does also show a lower error for the low-frequency wave height.

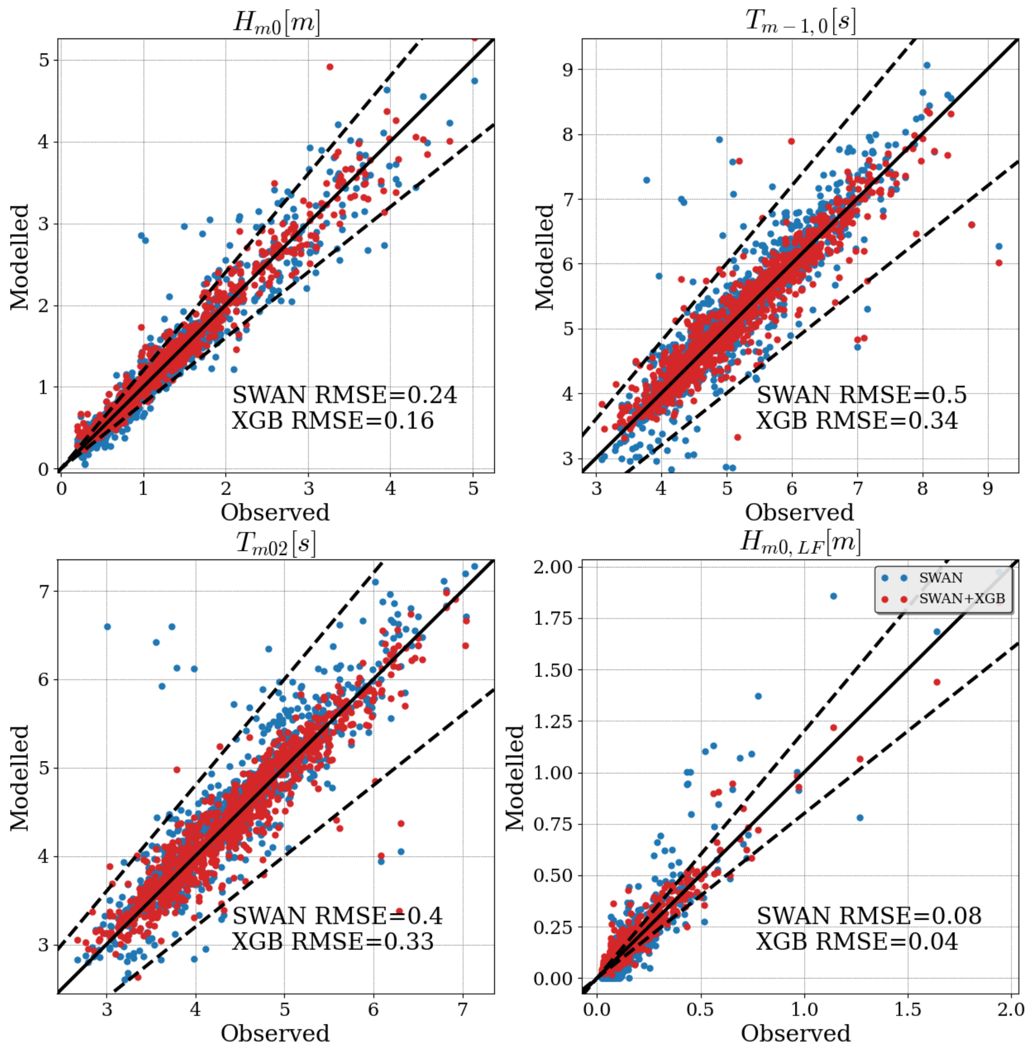

Figure 5. Scatter plot of the total wave height, spectral period, mean absolute wave period and low-frequency wave height for the location Europlatform. The solid black line represents perfect agreement and the dashed lines show the 20% deviation.

4. Discussion

It is shown that the error in energy density and directions reduces after applying the correction predicted by the data-driven model, but no additional optimizations are applied. In this study, we assume a set of input features in the data-driven model to predict the correction. In future research, the data-driven model could be further optimized by varying the input features. This optimization will be a balance between the available data limiting the size of the training dataset and the additional physical information as input. It could also be questioned whether the different locations should be part of one model or whether a single model for each location should be trained. When the patterns in the input variables required for the correction are not very similar for the different locations, it could beneficial to train a model for each location. Furthermore, a set of hyperparameters for the data-driven model are applied. By varying these hyperparameters, the training of the data-driven model could be improved, which could also lead to an improved correction.

5. Conclusion

A data-driven model has been trained to predict the correction for the energy density spectrum and directional spectrum for three locations in the North Sea obtained with the SWAN model. Trained on one year of observations and model results, the data-driven model turns out to significantly improve the SWAN forecasts. Both the energy density and the directions show a reduction in RMSE of up to 30% for the directions and 26 % for the energy density. Due to the short computational time of the data-driven model, this approach can easily be implemented in operational forecast systems.

Acknowledgements

We would like to thank Rijkswaterstaat WVL for providing the observed wave data and funding this study.

References

Booij, N., Holthuijsen, L. H., and Ris, R. C. 1997. The "SWAN" wave model for shallow water. In *Coastal Engineering* 1996 (pp. 668-676).

Callens, A., Morichon, D., Abadie, S., Delpey, M., & Liquet, B. (2020). Using Random forest and Gradient boosting trees to improve wave forecast at a specific location. Applied Ocean Research, 104, 102339.

Chen, T., and Guestrin, C. 2016, Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Deltares, 2018. SWAN-Kuststrook dd November 14th, 2018. Ref 12202221-005-ZKS-0005.

Gautier, C, and Caires, S (2015). Operational wave forecasts in the southern North Sea. In *proceedings of the 36th IAHR World Congress.*

Thomas, T. J., and Dwarakish, G. (2015). Numerical wave modelling–A review. In *International Conference On Water Resources, Coastal and Ocean Engineering* (ICWRCOE 2015). Aquatic Procedia (Vol. 4, pp. 443-448).

Van Dongeren, A., Van der Westhuysen, A., Groeneweg, J., Van Vledder, G., Lansen, J., Smale, A., ... & Wenneker, I. (2011). Spectral wave modelling in tidal inlet seas: Results from the SBW Wadden Sea project. *Coastal Engineering Proceedings*, (32), 44-44.