Chlorophyll Level Estimation
Based On Remote Sensing Reflectance

The goal is to develop a prediction model that can estimate the chlorophyll (CHL) level based on satellite reflectance data. The remote sensing reflectance (RRS) spectrum at different wavelengths is used as input data. The wavelengths used to record the RRS spectra are listed in Table 1.

λ (nm)	400	412.5	442.5	490	510	560	620	665	673.75	681.25	708.75	753.75	761.25	764.375	767.5	778.75
Label	WL400	WL412	WL442	WL490	WL510	WL560	WL620	WL665	WL673	WL681	WL708	WL753	WL761	WL764	WL767	WL778

Table 1: The wavelengths used to record the RRS spectra.

The CHL distribution for the training dataset and the validation dataset is shown in Figure 4 and Figure 5. The RRS spectra for the training dataset and the validation dataset is shown in Figure 8 and Figure 10, while their correlation with the CHL level is shown in Figure 9 and Figure 11. Since the CHL distribution is highly peaked towards lower values, a downsampling process is performed on the training dataset.

The sampling process consists of randomly selecting the same number of samples for each CHL bin. The chosen CHL bin size is 2, while the number of samples per bin is set to 1566 to maximize the retention of the statistics at higher CHL values. Figure 6 shows the CHL distribution for the sampled training dataset. The RRS spectra and their correlation with the CHL level for the sampled training dataset is shown in Figure 12 and Figure 13.

To reduce the skewness of the input data, further scaling and transformations are applied. The RRS spectra are transformed as:

$$X' = log(100 * X)$$

while the CHL level is transformed as:

$$Y' = log(\frac{Y}{20})$$

All these transformations are applied to the training, validation and test datasets. The CHL level for the processed training dataset is shown in Figure 7, while the RRS spectra and their correlation with the CHL level are shown in Figure 14 and Figure 15.

To boost the prediction power of the models under investigation, the input data is augmented by ratios of RRS spectra. All possible combinations of higher wavelength spectra over lower wavelength spectra are added as input. In total, 120 ratios are added for each dataset. Examples of ratios of RRS spectra and their correlation withe CHL level are shown in Figure 16 and Figure 17.

The first model under investigation is a random forest regression model using scikit-learn. For this model, the following hyper-parameters are considered:

- the number of trees in the forest (N_TREES)
- the minimum number of samples required to be at a leaf node (MIN_N_LEAF)
- the number of samples to draw from X to train each base estimator (MAX_SAMPLES)

To find a good hyper-parameter combination, a rough grid search is performed. Table 2 lists the hyper-parameter values under investigation. The best prediction is obtained for:

- N_TREES = 64
- MIN_N_LEAF = 4
- MAX_SAMPLES = 0.5

	V1	V2	V3
N_TREES	32	**64**	128
MIN_N_LEAF	2	**4**	8
MAX_SAMPLES	**0.5**	0.7	0.9

Table 2: The hyper-parameter search space for the random forest model.

Figure 1 shows the predicted CHL level versus the true CHL level for the validation dataset with the corresponding MSE value of 48.23.

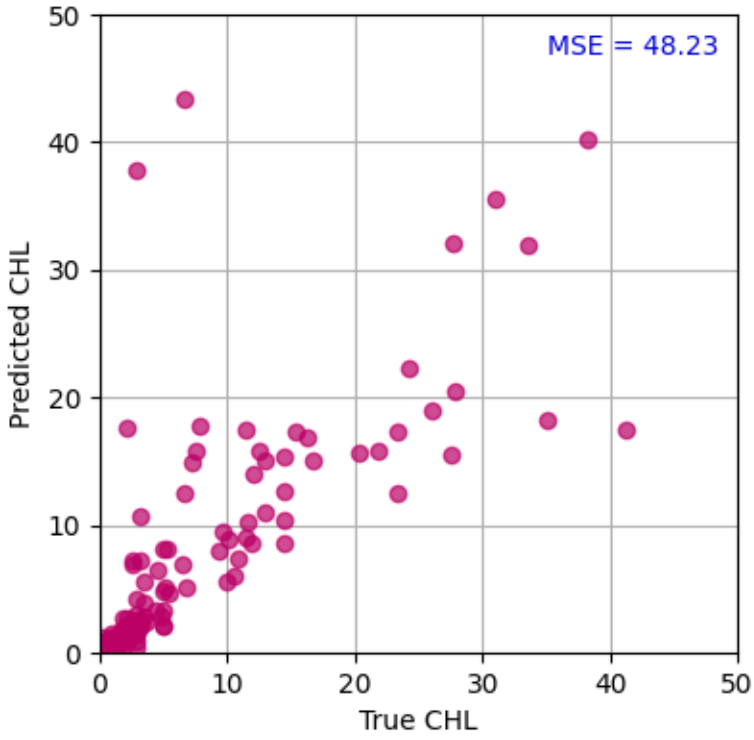

Figure 1: The predicted versus the true chlorophyll level using the random forest model.

The uncertainties associated with the model predictions are estimated using a bootstrapping approach. For the random forest, the bootstrapping technique is part of the model, namely each tree in the forest runs on a bootstrap sample. Therefore, the predicted value represents the mean of all the tree predictions, while the uncertainty represents the standard deviation of the tree predictions. Other sources of uncertainty would be related to the measured data, in which case an error propagation approach would be employed.

Another model under investigation is a fully connected neural network model using TensorFlow. The basic design of the model consists of:

- one input layer
- one dense layer with a RELU activation function and a L2 regularizer
- one dropout layer
- one dense layer with a RELU activation function and a L2 regularizer
- one dropout layer
- one dense layer with one node

The neural network model is trained using the:

- RMSProp optimizer
- mean squared error loss function
- batch size of 30 for 120 epochs

For this model, the following hyper-parameters are considered:

- number of nodes in the hidden layers (UNITS)
- regularization rate (REG_RATE)
- dropout rate (DROP_RATE)
- learning rate (LEARN_RATE)

To find a good hyper-parameter combination, a rough grid search is performed. Table 3 lists the hyper-parameter values under investigation. The best prediction is obtained for:

- UNITS = 256
- REG_RATE = 0.01
- DROP_RATE = 0.1
- LEARN_RATE = 0.00001

	V1	V2	V3
UNITS	64	128	**256**
REG_RATE	0.1	**0.01**	0.001
DROP_RATE	**0.1**	0.3	0.5
LEARN_RATE	0.001	0.0001	**0.00001**

Table 3: The hyper-parameter search space for the neural network model.

Based on the results of the feature importance for the random forest model and to boost the stability of the neural network model, only ratios of the RRS spectra are used for the neural network model.

Figure 2 shows the training and validation loss functions, while Figure 3 shows the predicted CHL level versus the true CHL level for the validation dataset with the corresponding MSE value of 54.58.

Given the current results, the random forest model is used on the test dataset to predict the CHL level and their corresponding uncertainties.

The next steps would be to:

- include measurement data in the training process
- do more or different data processing (scaling, normalization, transformation,...)
- fine tune the hyper-parameters of the selected models

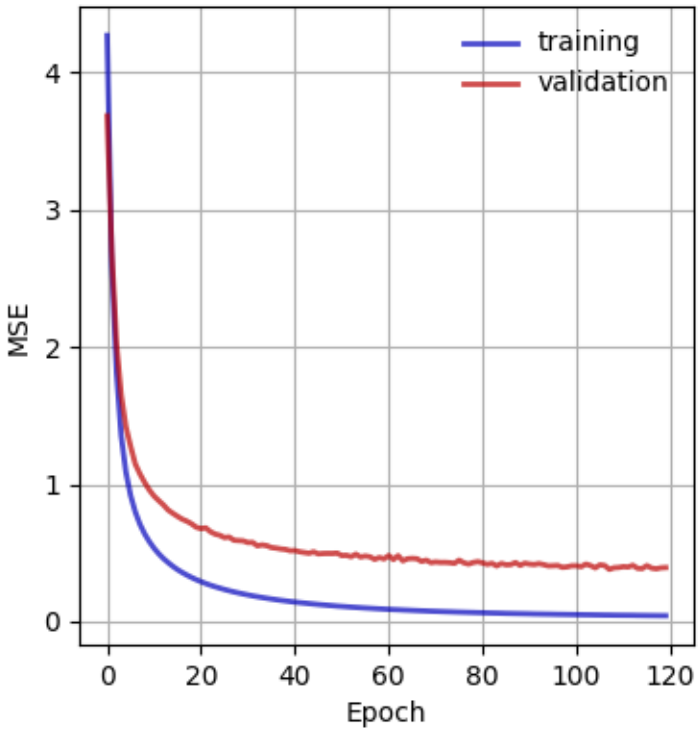

Figure 2: The training and validation loss functions for the neural network model.

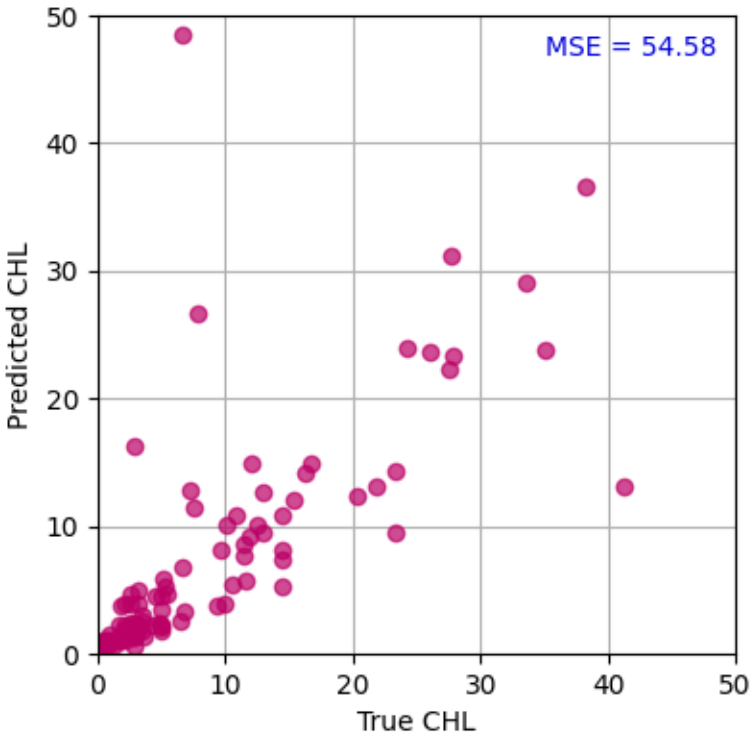

Figure 3: The predicted versus the true chlorophyll level using the neural network model.

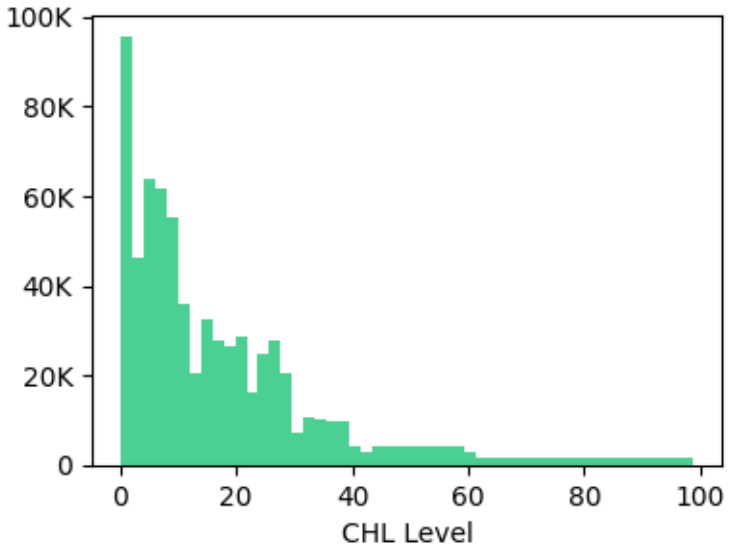

Figure 4: The chlorophyll level using the entire training data.

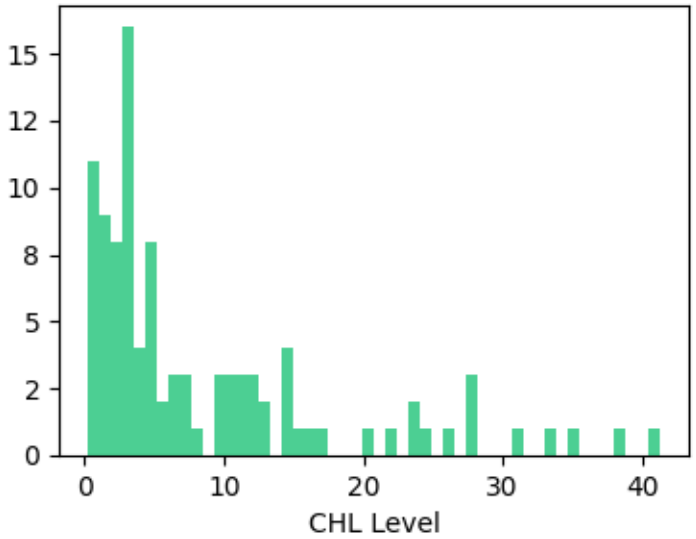

Figure 5: The chlorophyll level using the validation data.

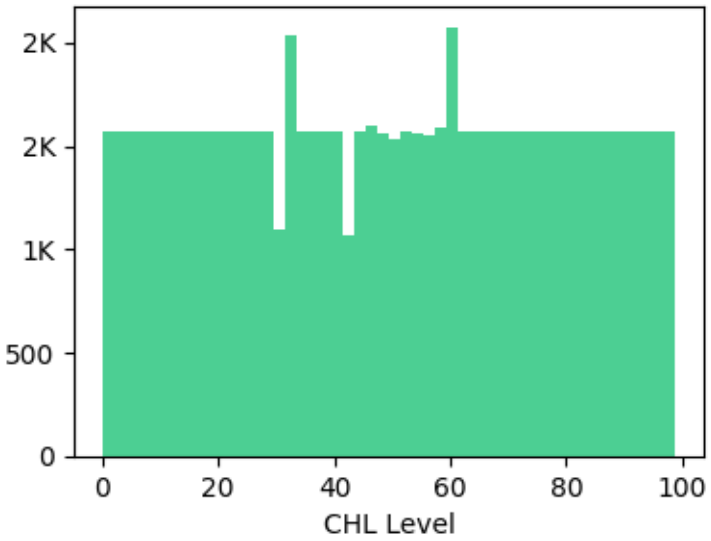

Figure 6: The chlorophyll level using the sampled training data.

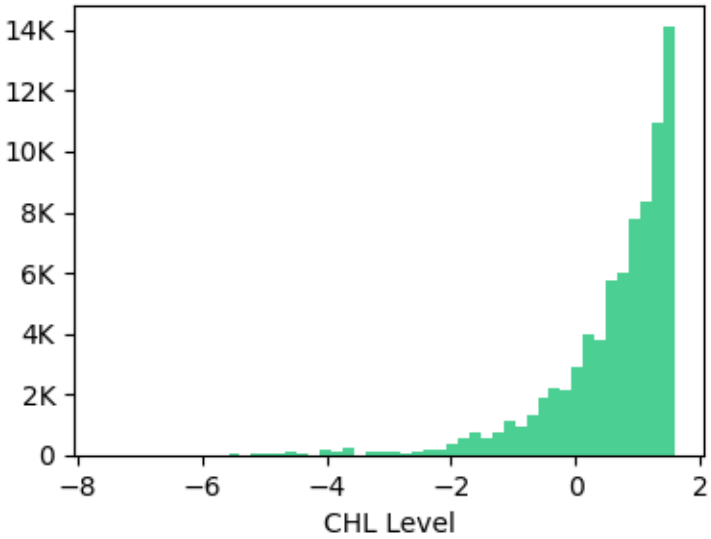

Figure 7: The chlorophyll level using the processed training data.

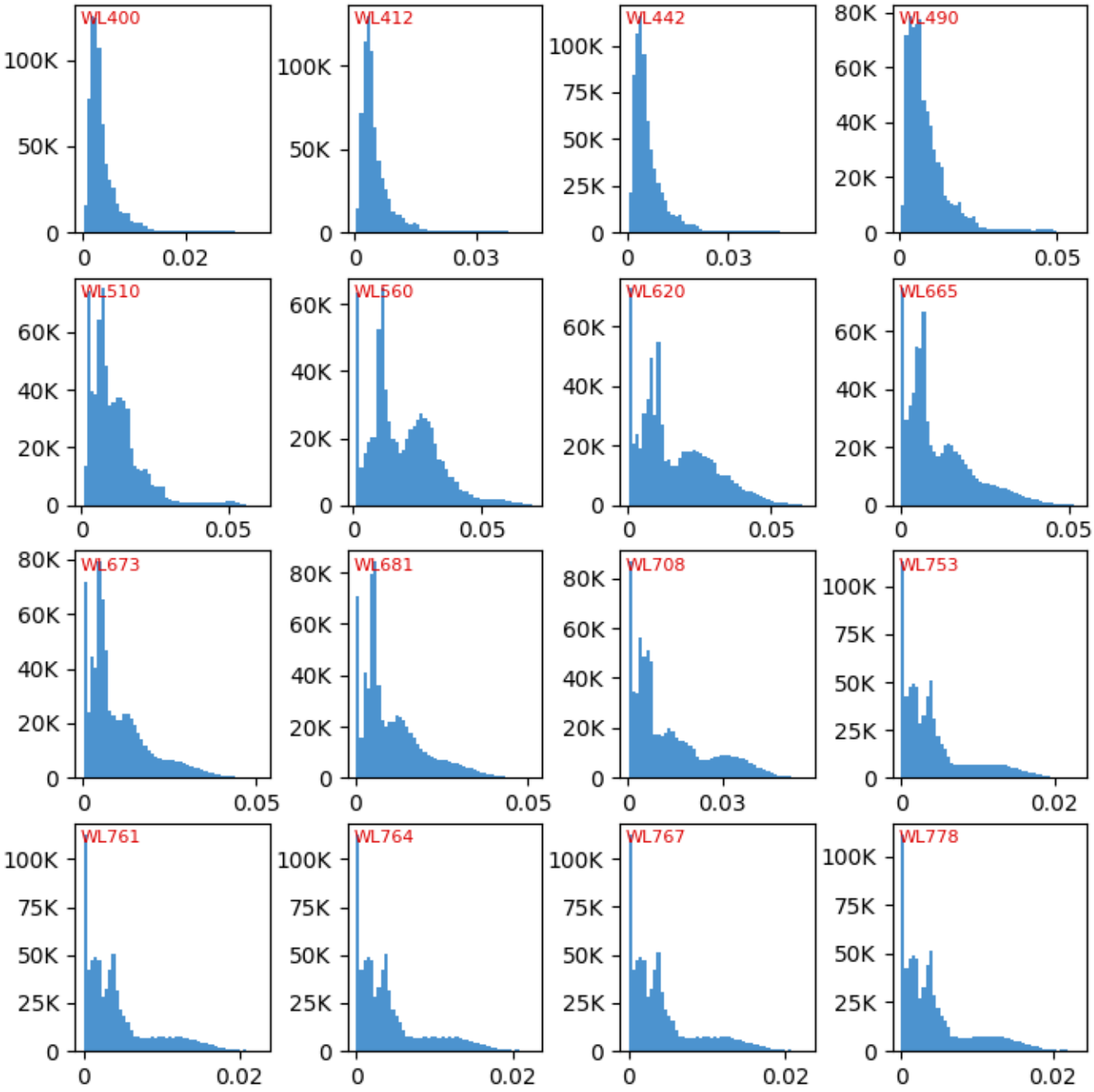

Figure 8: The RRS spectra using the entire training data.

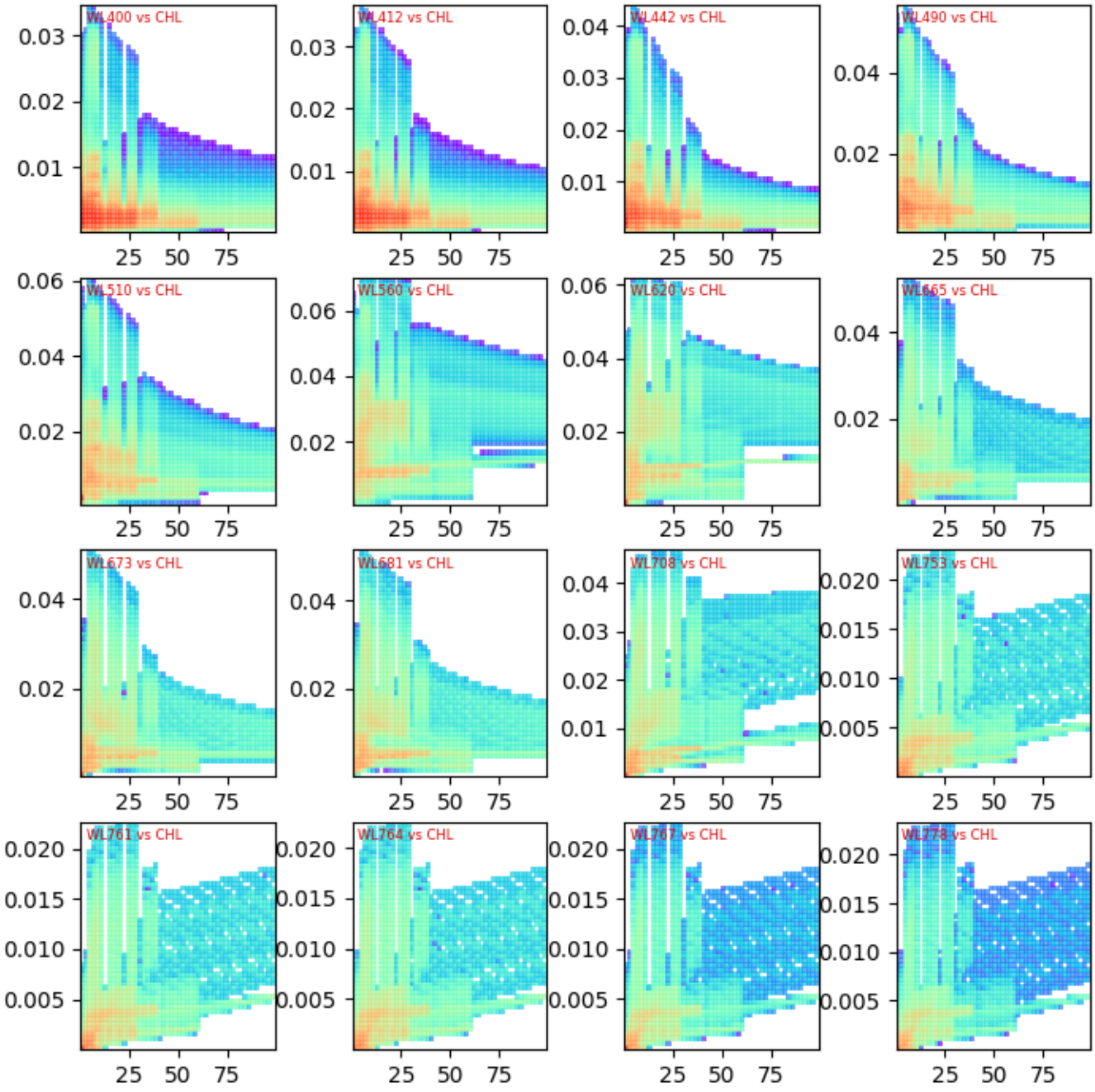

Figure 9: The RRS spectra versus the chlorophyll level using the entire training data.

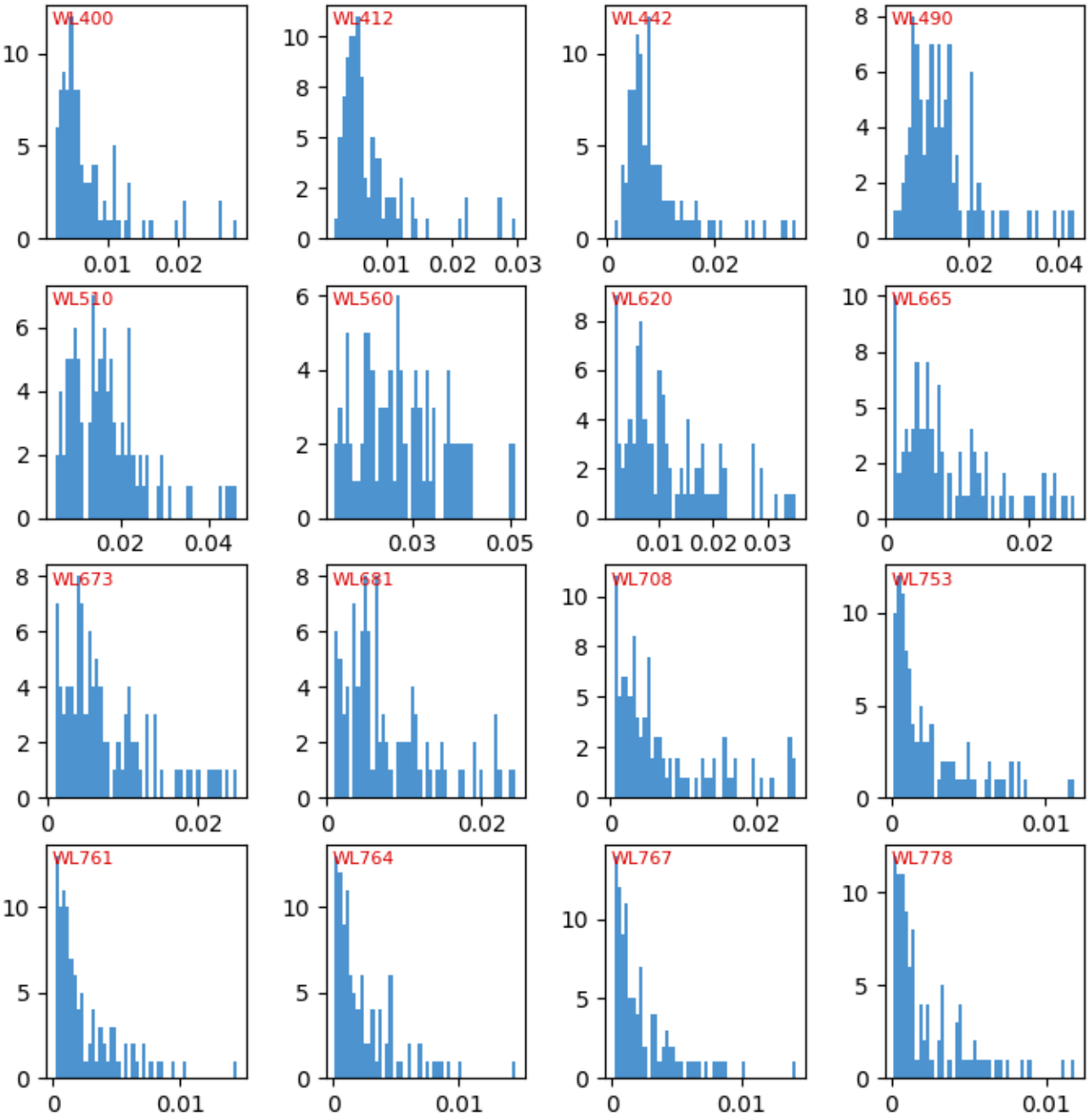

Figure 10: The RRS spectra using the validation data.

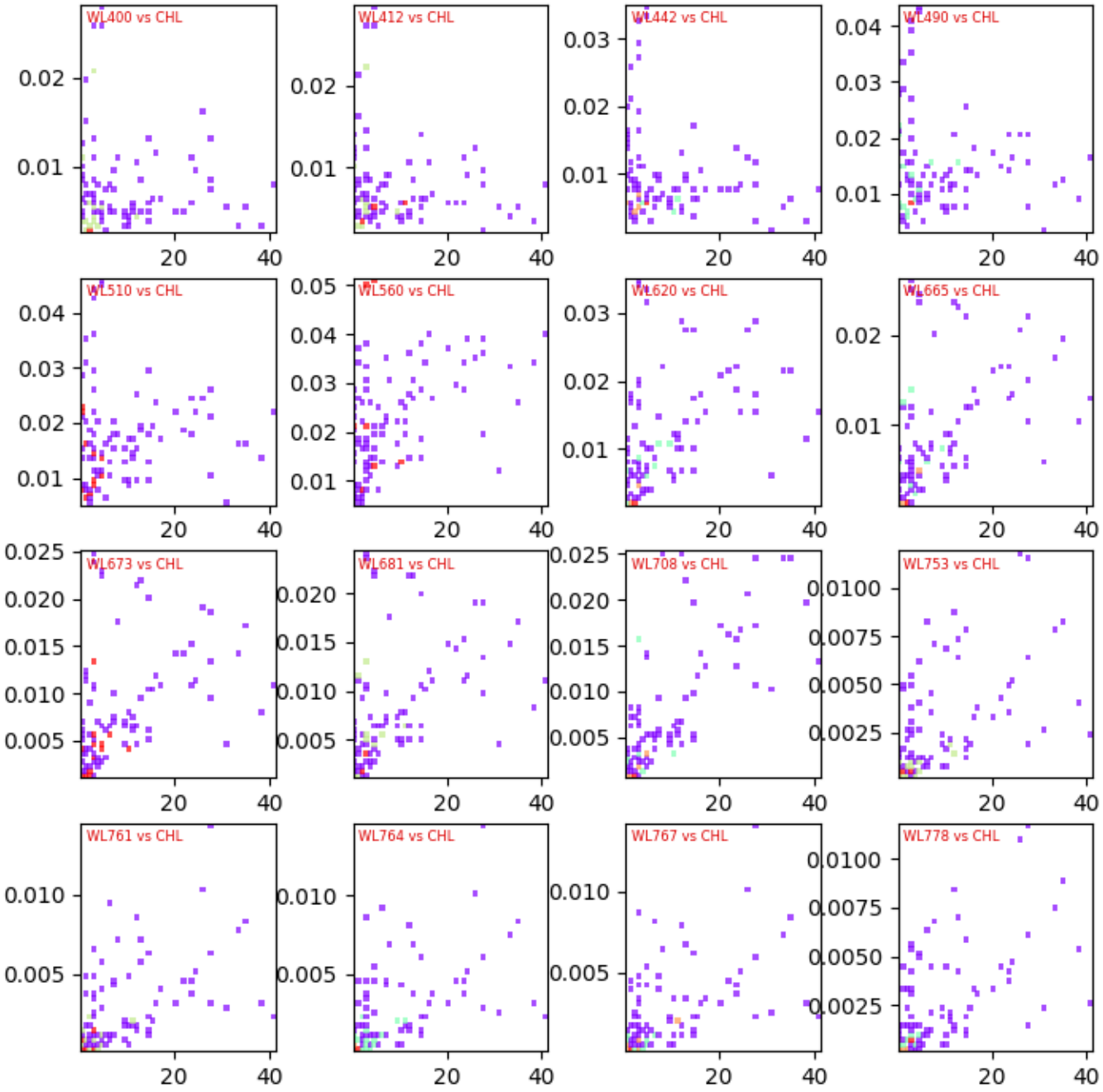

Figure 11: The RRS spectra versus the chlorophyll level using the validation data.

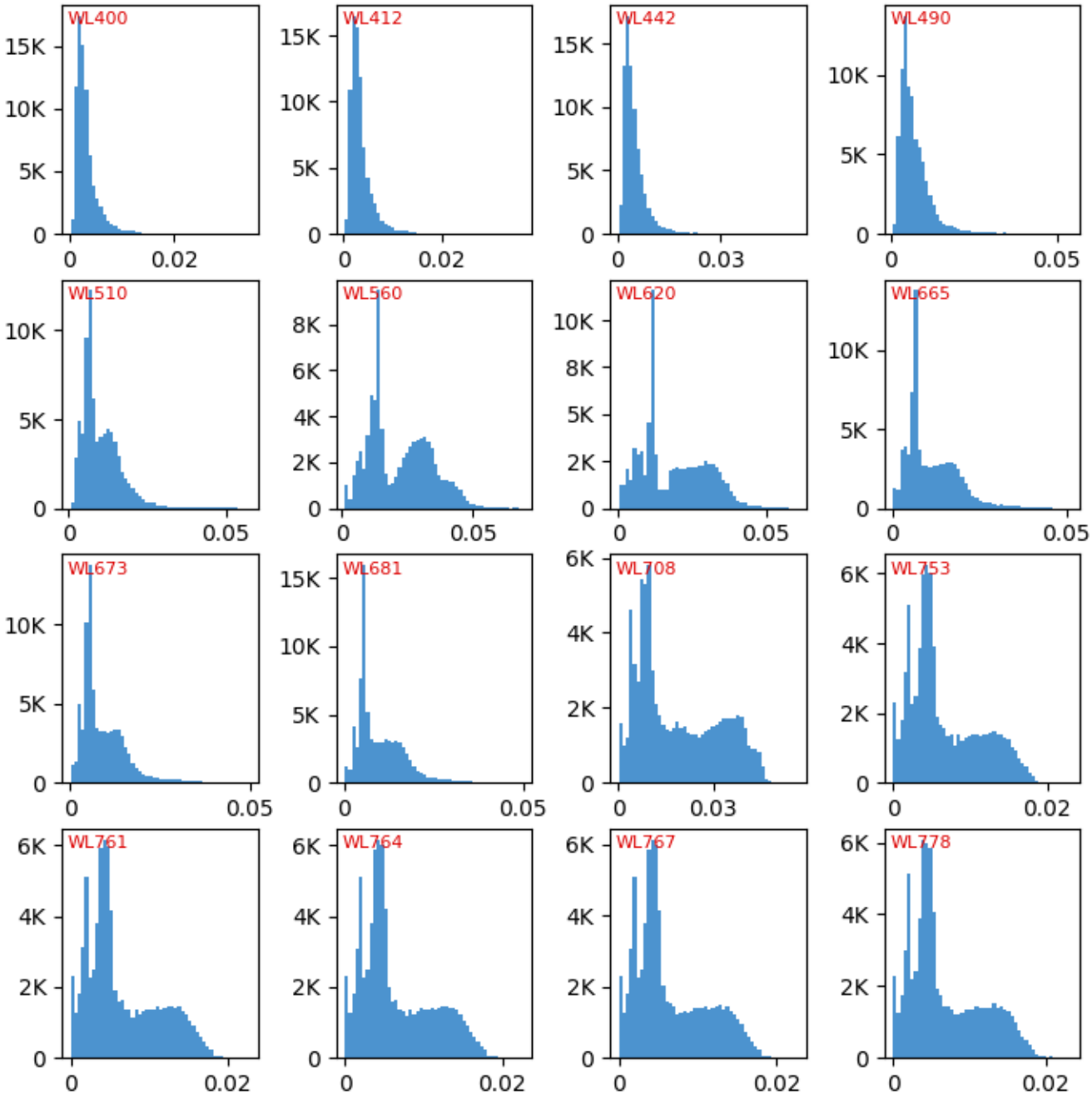

Figure 12: The RRS spectra using the sampled training data.

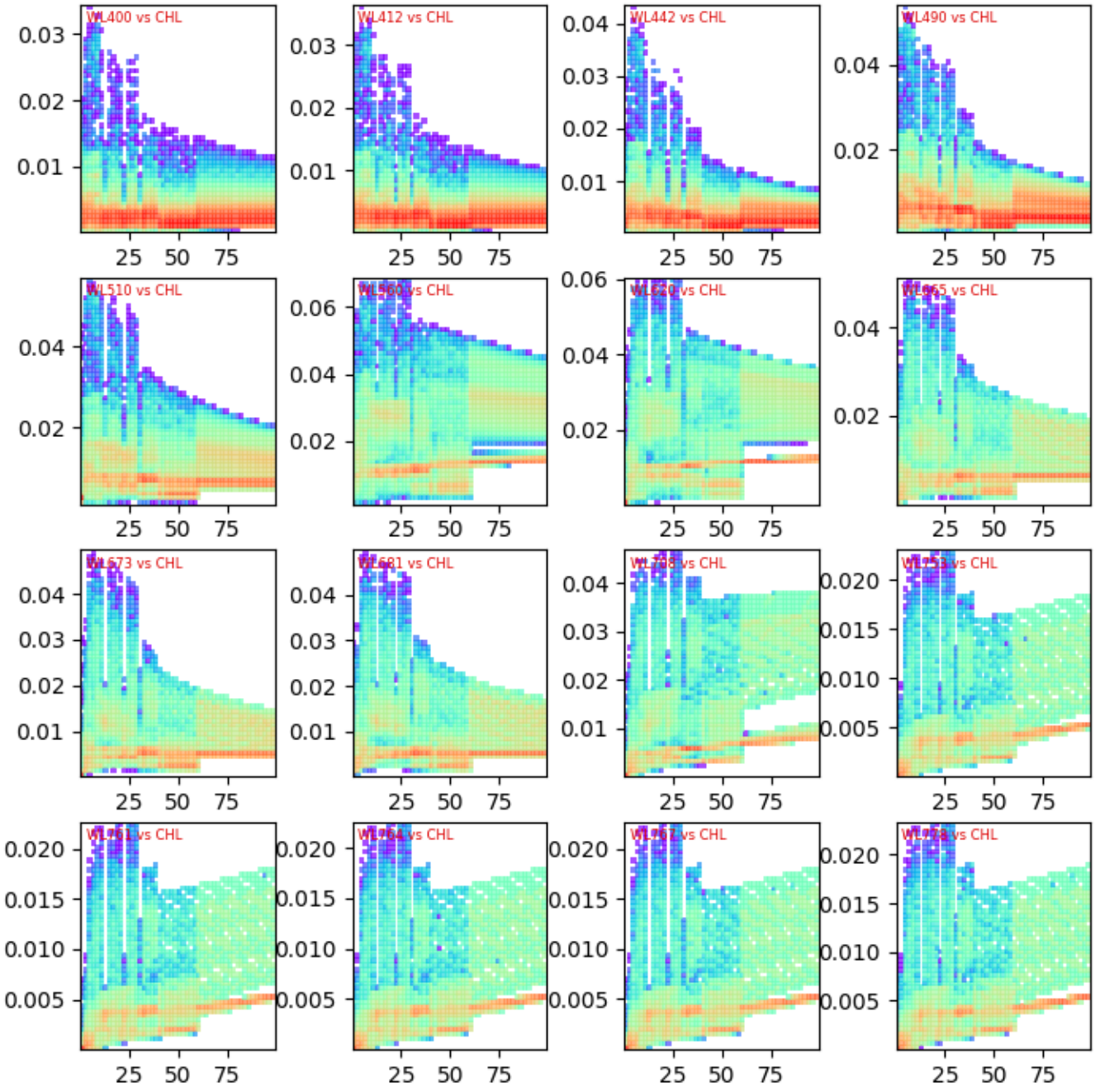

Figure 13: The RRS spectra versus the chlorophyll level using the sampled training data.

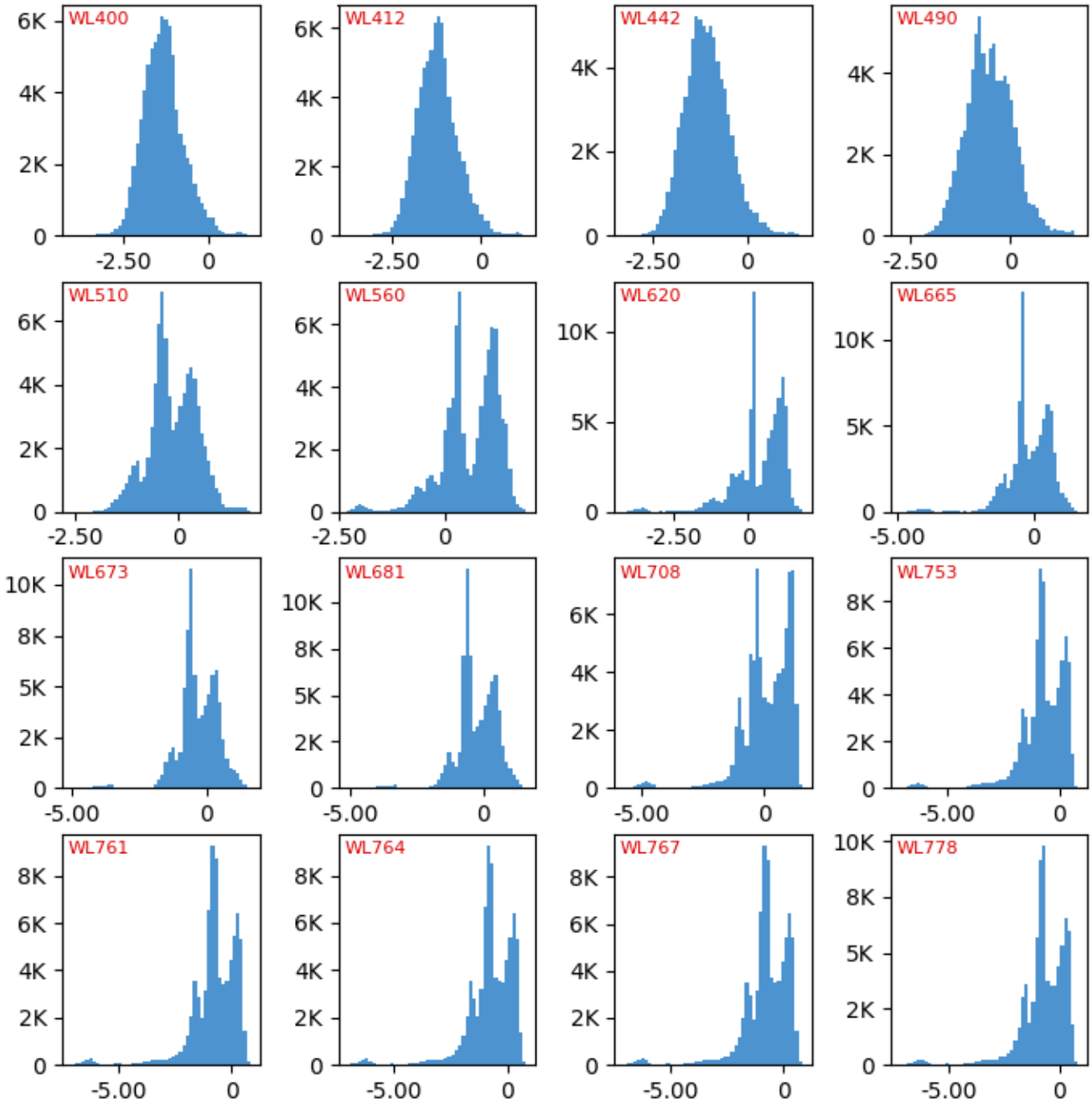

Figure 14: The RRS spectra using the processed training data.

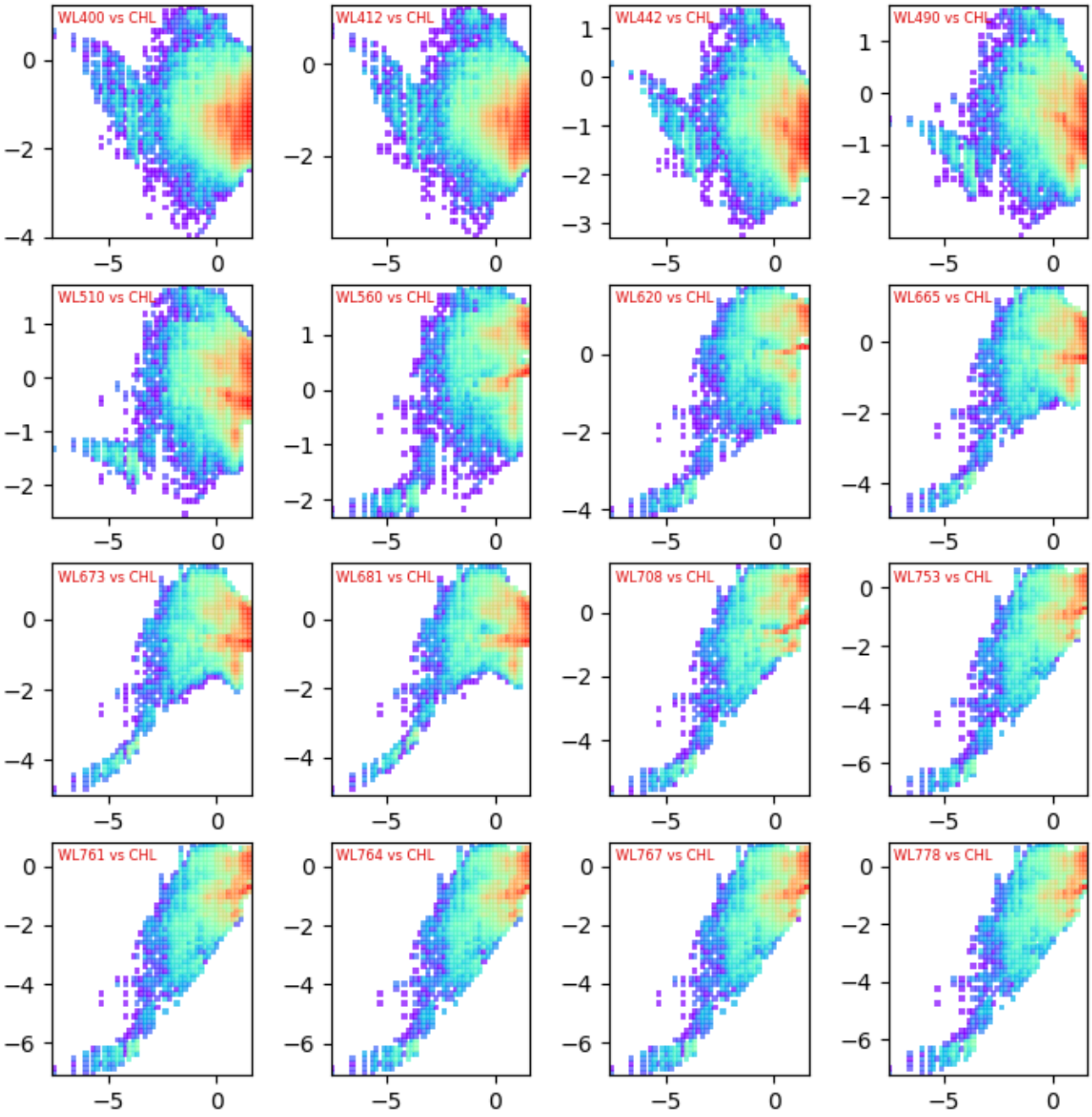

Figure 15: The RRS spectra versus the chlorophyll level using the processed training data.

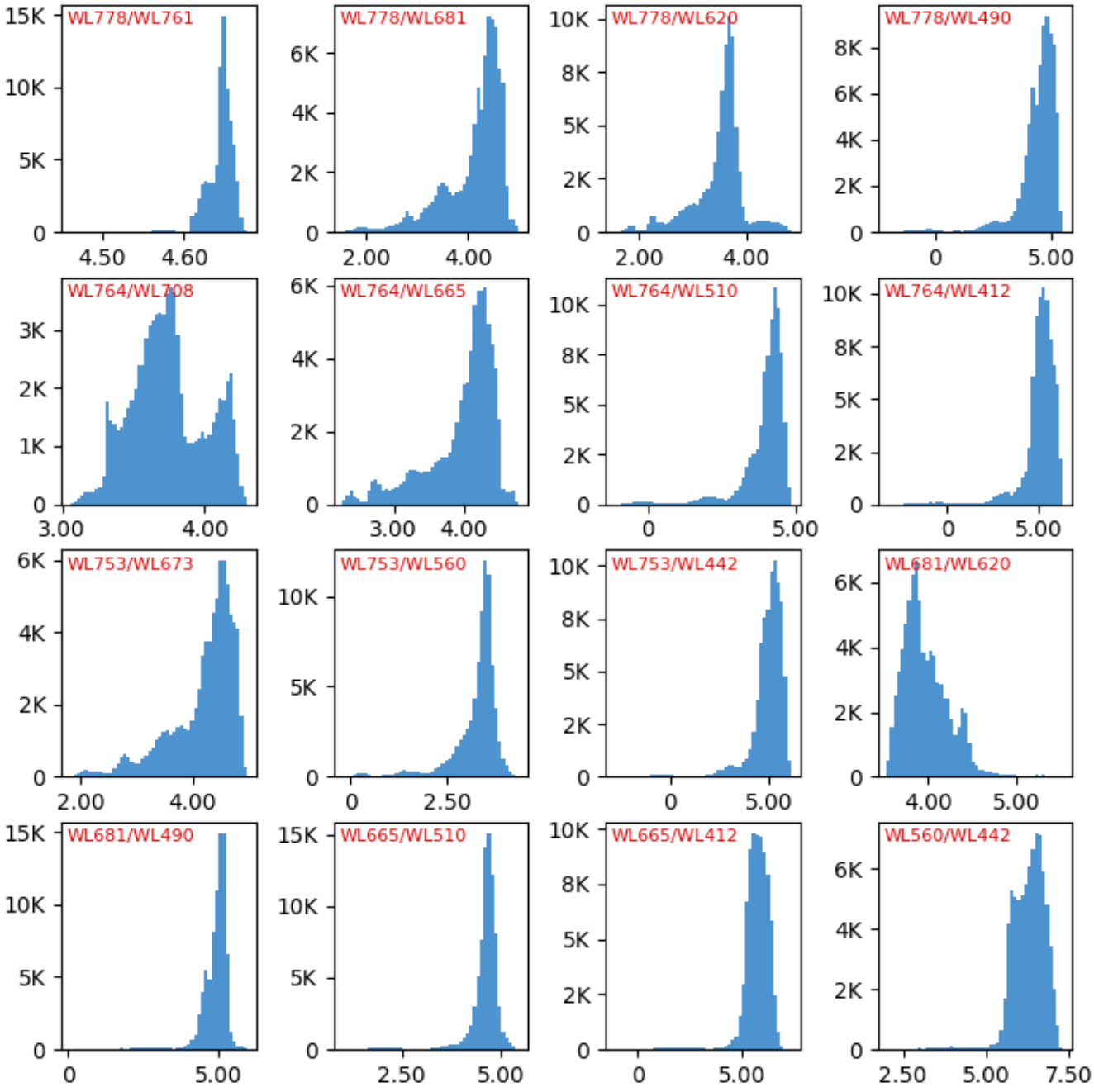

Figure 16: Ratios of RRS spectra using the processed training data.

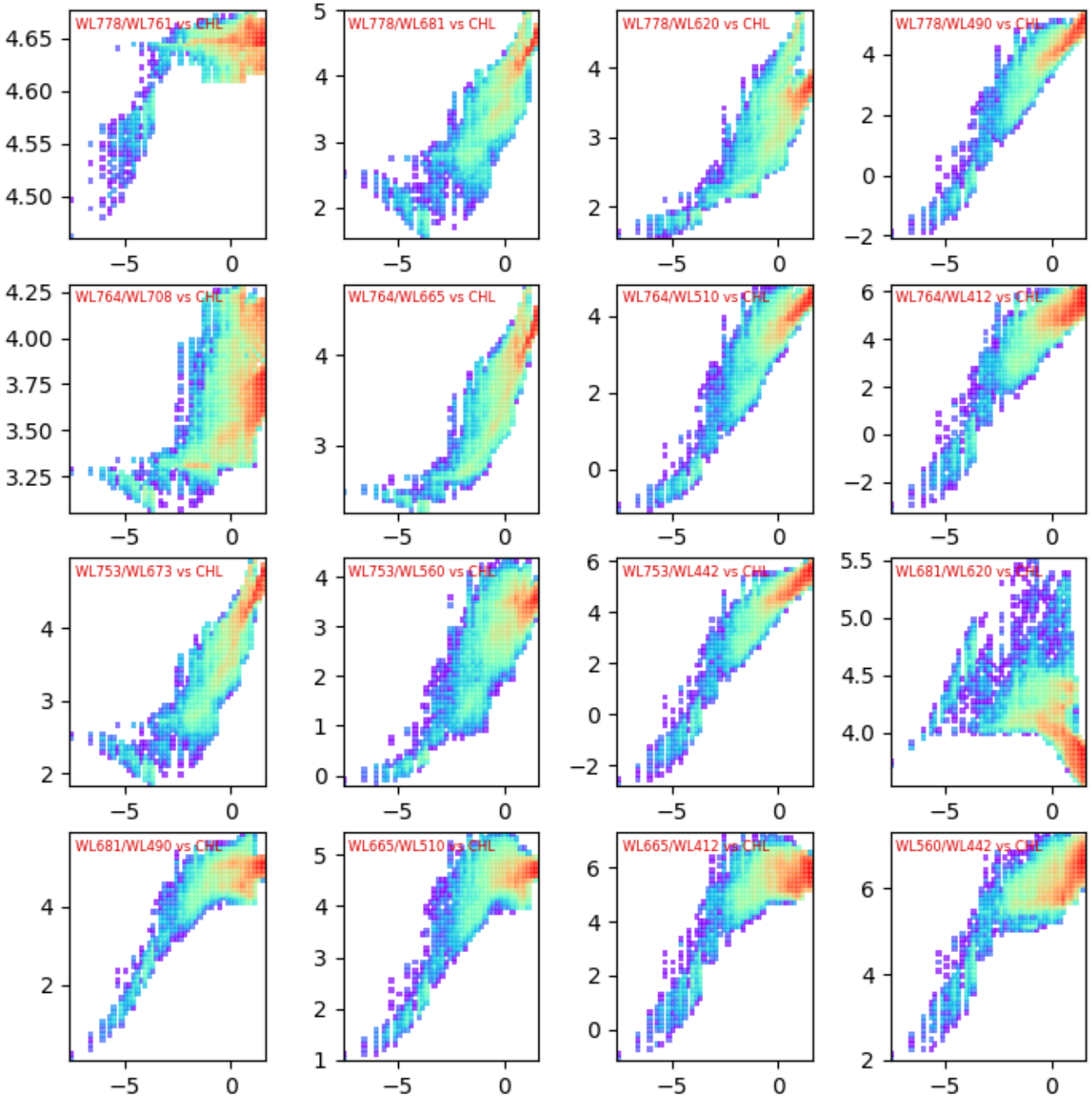

Figure 17: Ratios of RRS spectra versus the chlorophyll level using the processed training data.