To forecast California’s green energy production, 6 recurrent neural networks with linear hidden layers were used. The model assumes that each category of production, biogas, biomass, geothermal, small hydro, solar, and wind, are independent of each other. While this is untrue in the real world, as each category’s production affects how much each of the others can produce. For example, if the California government decides that there can only be 10,000MWh produced, and solar has already produced 8,000MWh. It would then be impossible to produce 5,000MWh of geothermal energy. In the model setting, creating independent neural nets reduces the output dimension, reduces the training time, and simplifies forecasts.

A recurrent neural network is one where previous information is preserved. At each forward pass, the network takes in new data as well as information from previous inputs. RNN’s are well suited for natural language processing, where the order of letters or words matter, and time series. With our models, we took advantage of latter. The network is structured with two linear hidden layers, leaky ReLU activation functions, batch normalization after each linear layer, and one linear output layer. During training, we used Adam to update gradients, mean squared error as a loss function, and trained for 200 epochs. To generate features to regress on, we used 100 lags, or the 100 previous datapoints in the time series, as well as basic, one-hot encoded features like day of the week, month, year, and so on. We then scaled the lags with sklearn’s robust scaler, which is similar to subtracting the mean and divided by standard deviation but less affected by outliers.

We reserved the last 20% of the data, roughly two years from the beginning of 2016 until 2018, as our test set. For each model, the test set had a RMSE of roughly 1500-2000. While this sounds like a lot, the data is incredibly noisy, and using 100 lags significantly smoothed out the noise. We also used several regularization techniques (batch norm, dropout) to avoid overfitting, thus increasing the loss.

Lastly, we extrapolated the result of the RNN’s to predict the years following the data set, from January of 2018 until January of 2021. In general, the model predicts reductions in wind power and biogas, steady production in geothermal and biomass, and increases in solar and hydro power.

One issue with the models is the lack of features. Wind and solar in particular are dependent on weather conditions. Adding in information about weather, demand for each type of production, and funding for renewable energy production may improve performance. Vanilla RNN’s, like the ones we used, “lose” information as time goes on. There is an inherent recency bias built into the network. This can fixed with attention or using other structures like an LSTM. Also, using 100 lags may be an inappropriate amount, and a lower number may better simulate the seasonality of green energy production.