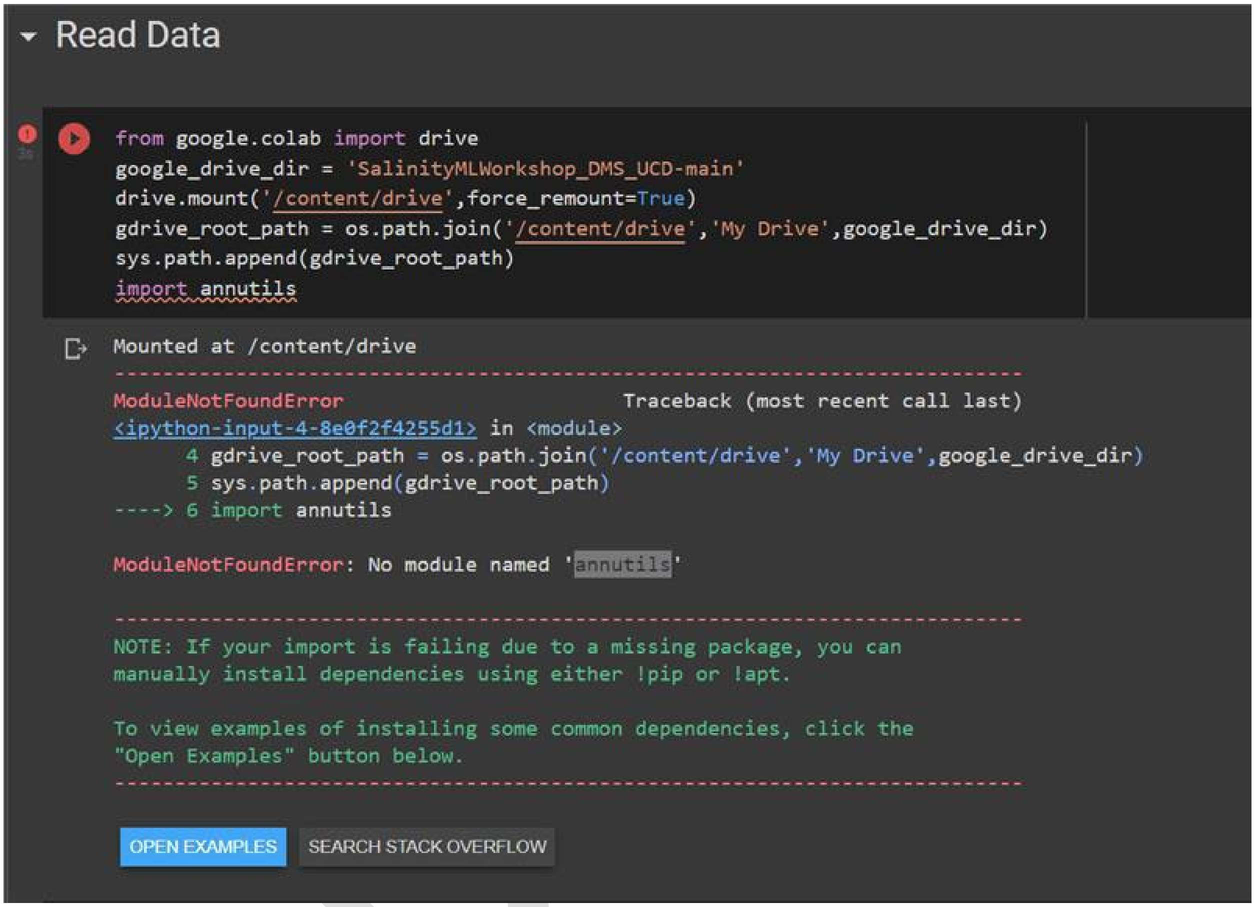
Questions and Answers from Delta Flow-Salinity Modeling Using Machine Learning Workshop

Setup and Intro

Q: I was trying to do the setup this morning and after opening Colab and running the first script (`Colab_Train_ANN_on_Augmented_Dataset.ipynb`), I get the following error:



Read Data

```
from google.colab import drive
google_drive_dir = 'SalinityMLWorkshop_DMS_UCD-main'
drive.mount('/content/drive',force_remount=True)
gdrive_root_path = os.path.join('/content/drive','My Drive',google_drive_dir)
sys.path.append(gdrive_root_path)
import annutils
```

```
Mounted at /content/drive
---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last)
<ipython-input-4-8e0f2f4255d1> in <module>
      4 gdrive_root_path = os.path.join('/content/drive','My Drive',google_drive_dir)
      5 sys.path.append(gdrive_root_path)
----> 6 import annutils

ModuleNotFoundError: No module named 'annutils'

---------------------------------------------------------------------------
NOTE: If your import is failing due to a missing package, you can
manually install dependencies using either !pip or !apt.

To view examples of installing some common dependencies, click the
"Open Examples" button below.
---------------------------------------------------------------------------
```

- A: Please make sure that the folder in your google drive is named: *"SalinityMLWorkshop_DMS_UCD-main".* Note: if you used git clone, the filename may not have "-main" at the end.
- Post-Workshop Response: As currently configured, the Google Colab option does not use relative paths, so please make sure that (1) the folder you upload is in the top-level of your Google Drive and (2) the name of the folder is consistent with the path in the Jupyter Notebook *.ipynb* file (the default folder name is *SalinityMLWorkshop_DMS_UCD-main*). Alternatively, you may change the pathname within the code to point to the directory of your choice.
- Comment: I added a '/' at the end of the directory in the second line in that box and fixed the problem.

  So, the second line should read:

  google_drive_dir = 'SalinityMLWorkshop_DMS_UCD-main/'

Module 2: Dashboard (Raymond Hoang)

Questions from the Teams Chat

Q: Are daily input values interpolated?

- A: Monthly scalars applied uniformly to every day of that month.
- A: All ML options trained separately with same data sets.
- A: the emulator is a daily model, values vary day by day and are scaled by the factor from the slider.

Q: Is there a preferred browser for the dashboard?

- A: No, any modern browser should work. Internet Explorer is not recommended though. Mobile browsers should work as well.

Q: Can the dashboard handle multiple years simulation?

- A: The dashboard only simulates one year at a time now (i.e., no antecedent conditions if you look at one year to the next). You can pick any water year from 1991 to 2018.

Q: Are there any future plans to have the output be map based?

- A: Yes, that is on the to-do-list.

Q: if you envision the next version of this emulator tool, what does the future study/improvements entail?

- A: more sophisticated changes, more options for advanced users, more interactivity, run your own inputs, map view/spatial output.

Q: How long did it take to develop this viewer?

- A: On and off for 5-6 months, probably a couple of months for someone new to dashboarding learning Azure and dashboarding with Python. Linking ML was very straight forward, but making it look good and responsive was hard part.

Q: Is it possible to add a feature to change the physical configuration in the Delta? (e.g., add a barrier at False River?)

- A: Yes, it is possible. To do this, we would have to feed more data to ML during training (e.g., DSM2 scenarios with and without barriers).

DRAFT

Q: Will slides be shared?

- A: The slides was uploaded to the GitHub repository https://github.com/CADWRDeltaModeling/SalinityMLWorkshop_DMS_UCD (in the folder /docs).

Q: Was it hard to get DTS to let you use Azure?

- A: Nicky Sandhu (prabhjot.sandhu@water.ca.gov) can help you address technical/set up questions regarding Azure.

Q: Have you run statistics for goodness of fit between DSM2 and ML?

- A: Yes, those metrics are presented in the publications detailing the machine learning models: https://doi.org/10.3390/w14223628 and https://doi.org/10.3390/w14132030.

In-Person Questions

Q: Which architecture to use and biases?

- A: The six different architectures play a different role. MLP a basic feed-forward; is Res-net has a shortcut path and performs slightly better than MLP, four recurrent-based networks perform best in this study.

Comment: are there plans to explore use of similar tools for various parameters (e.g., velocity)?

- A: Yes, can be done. For example, in a separate study we are looking at ion concentrations with the Delta.

Q: Does it look at different Delta cross channel operations?

- A: Dashboard is showing Delta Cross Channel gate historical operations. Changing operations are possible, but that functionality is not incorporated into this iteration of the Dashboard.

Q: With more training and more data can get near 100% accuracy?

- A: Additional training can result in incrementally higher accuracy but need to be cautious of overfitting.

Q: Link to the journal article published in 2021 in WRPM?

- A: https://ascelibrary.org/doi/abs/10.1061/%28ASCE%29WR.1943-5452.0001445 (also available under /docs in the workshop GitHub repository)

Comment: each "stutter" is the emulator re-evaluating the new inputs, based on the user-input slider value.

DRAFT

Module 3: Flow Salinity Modeling using ANNs (Siyu Qi)

Q: Results are pickled. How to unpickle?

A: need script to unpickle results

Q: what do pickle and unpickle mean?

A: “Pickling” is a method to compress and package Python objects; “Unpickling” is extracting the archive. See pickle — Python object serialization — Python 3.11.1 documentation.

Q: With sigmoid activation function in MLP architecture, have you observed vanishing gradient problem?

- A: Yes, in beginning we encountered the vanishing gradient problem often. To solve this problem, we replaced output layer with Rectified Linear Unit (ReLU) function.

Q: To change delta cross channel operations, do you use only new DSM2 results to train ANN or do you add it to the previous data.

- A: this is data augmentation, we did train ANN using different DSM2 simulations, use old and new DSM2 results for training. This will be discussed further in an off-line meeting.

Q: In general, what is the most important factor to consider when deciding on the architecture? (e.g., the physics to be emulated, the nature of the input data, etc.)

- A: The six different architectures play a different role. MLP a basic feed-forward; is Res-net has a shortcut path and performs slightly better than MLP, four recurrent-based networks perform best in this study. Generally, the number of layers, number of neurons on each layer, activation function selection and training datasets are important factors that will impact the performance.

Q: Where in the script do we conduct parameter sweeping?

- A: Parameter sweeping was done in earlier phases of this study, where we selected the optimal selection of hyperparameters, so it is not done explicitly in the script.

Q: What are considerations for “performance”?

- A: training time and test speed, plus four more evaluation metrics including correlation, percent bias, RSR, and NSE.

DRAFT

Q: How do we split training test splits?

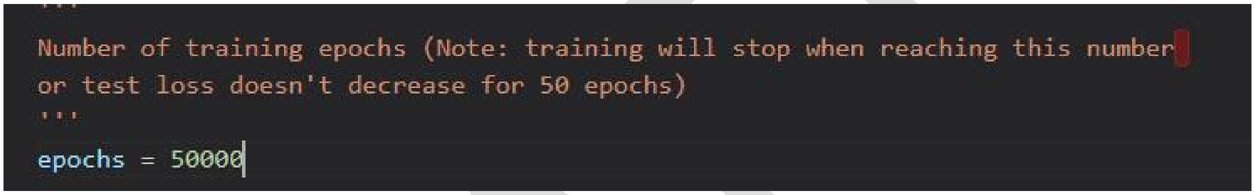
- A: chronologically 70% training 30% testing, train 8 hydro related training sets, test on 4 gate sets, larger training set gives better results, but need to have testing set not be too small. 70/30 and 80/20 are popular splits, exploratory studies 70/30 gave desired results and fair test set.

Q: If change hydro parameters do you need to rerun?

- A: Yes

Q: The simulations generally under-predicts the peak of the target. Is that because this is an example run and or is that a weakness even in the final trained model?

- A: Yes. If you increase epoch number in the first cell, you will have better results. For example:



```
'''
Number of training epochs (Note: training will stop when reaching this number
or test loss doesn't decrease for 50 epochs)
'''
epochs = 50000
```

- Post-Workshop Response: Increasing the epochs will improve performance. In workshop scripts, the epochs were reduced to decrease runtimes for demonstration purposes. For the published results, the training process stops either at epoch 5000 or when the test MSE does not decrease for 50 epochs.

Q: How big do the .pkl files get? Is it related to the epoch number?

- A: less than 20 KB.

Q: In the "Train Models" section, I get an error after epoch 200 that says the directory does not exist (running in Colab).

- A: Try rerunning the script or select runtime restart and run all.

Q: For the local scripts, is there an optimal hardware configuration if I wanted to run locally?

A: Tensorflow dropped support for GPU on Windows platform as of v.2.11 (which these scripts are based on). If you want to run on windows with GPU, then you need Windows subsystem for Linux. And run this workflow on the Linux subsystem; probably limited to running on CPU-only if using Windows or use a cloud-based service.

Q: how did you take care of the overfitting issue? did you use any validation test set?

- A: 70/30 does not overfit, also provide augmented data and transfer learning to avoid overfitting. See manuscripts https://doi.org/10.3390/w14223628 and https://doi.org/10.3390/w14132030.

DRAFT

Module 4: Physics Informed Neural Network (Dong Min Roh)

Q: Are the physics time varying? Is this a recursive approach?

- A: physics are time varying with time dimension.

Q: So, it is recursive?

- A: No.

Q: please post link to PINN survey review article Dong mentioned

- A: article was uploaded to /docs of the workshop GitHub repository.

Q: In the advection-dispersion equation, are we assuming that the cross-sectional area (A) and longitudinal dispersion factor (K) are constant?

- A: Yes, in this proof of concept, we are assuming that they are constant. In future models we would incorporate additional information to represent the A and K terms.

Q: If set advection-dispersion based loss function to zero would you effectively you get the same results as a pure ANN?

- A: Yes

Q: How are you solving L_F? Are you solving a separate physics-based model concurrently?

- A: With a pure ANN, only minimizes L_D when training, but with PINN we have an extra term L_F, which we train like a regular neural network (i.e., to get optimal theta which is the optimal weights and bias).

Q: Is the k-fold approach used minimize input data when input data is limited?

- A: The k-fold approach is used as a validation technique to see how well the model performs for the independent dataset.

Q: When dividing data into testing and training, choose best parameters, has K-fold been used before choosing final parameters?

- A: optimal parameters were found for each Kfold, then Kfold experiment was conducted

---

Q: is objective of PINNs to increase performance

- A: yes

DRAFT

Q: Can PINNS inform how well dispersion is estimated?

- A: Yes, that is known as an inverse problem. You can learn what salinity is to learn what outflow or various coefficients are. Could be a future direction.

Q: What if you turn up knob to 100% Physics based?

- A: Yes, but the performance would not be good – still need the ANN. PINN uses differential equation to try to model underlying physics as best as we can, still must use actual or simulated data sets, ANN is melding of the two. Note: the differential equation has simplifying assumptions for this proof-of-concept demonstration (e.g., dispersion coefficient and cross-sectional area are assumed constant)

DRAFT