25th International Congress on Modelling and Simulation, Darwin, NT, Australia, 9 to 14 July 2023
mssanz.org.au/modsim2023

Application of denoising diffusion models in the augmentation of a regional dataset of inherent optical properties and applications for remote sensing

Nathan Drayson [a], **Foivos Diakogiannis** [b], **Nagur Cherukuru** [a], **David Blondeau-Patissier** [c] **and Thomas Schroeder** [c]

[a] CSIRO Environment, Canberra, Australia; [b] CSIRO Data61, Perth, Australia
[c] CSIRO Environment, Brisbane, Australia
Email: nathan.drayson@csiro.au

Abstract: Optical water quality monitoring using remote sensing provides routine retrievals of in-water constituents such as total-suspended solids (TSS) and chlorophyll-a (CHL-a) concentration at a global scale. Increasingly, data driven methods such as deep learning are being used to derive in water constituents from optical signatures. These methods depend on large datasets to provide adequate training and validation data. As the acquisition of suitable training data is time consuming and expensive, model development has relied on international collaborations to supply sufficient observation volumes (Lehmann et al 2023). Such datasets are a valuable resource, however, by the nature of the collaborative effort provide data that is widely geographically dispersed and may not be suitable for region-specific algorithms.

Various methods have been implemented to address the issue of scarce training data (Cavalli, 2020). This study will present an evaluation of using denoising diffusion models, to generate synthetic inherent optical properties (IOPs) matched with TSS and CHL-a concentrations. We will also discuss the benefits of this augmentation method on retrieval accuracy of deep learning inversion models. The synthetic dataset is evaluated by examining the representation of IOPs to the training dataset. Of critical importance is the preservation of inter-feature statistical relationships such as between IOPs and constituent concentrations; and between the constituent concentrations themselves.

The 1D diffusion model was trained on a dataset comprised of 1049 complete sets of IOPs and constituent concentrations represented spectrally as an image. Each sample consisted of IOPs and constituent concentrations and included particulate absorption (a_p), absorption of coloured dissolved organic matter (a_{CDOM}), particulate scattering (b_p) and particulate backscattering (b_{bp}) TSS and CHL-a.

Similarities between feature distributions of the synthetic and measured datasets were evaluated using Jansen-Shannon (JS) distance, Kruskal-Wallis test and bootstrapped permutation tests. As b_p, b_{bp}, and a_{CDOM} can be represented by exponential decay functions, the coefficients describing these relationships were used to represent their feature distributions. To examine the representation of inter-feature relationships the bi-variate distributions of: $b_{bp}(550)$:TSS, $a_p(440)$:TSS, CHL-a:TSS, CHL:$a_{CDOM}(440)$, $a_p(676)/a_p(550)$:CHL/TSS were evaluated using JS distance and the statistical energy distance test (Aslan and Zech, 2005).

Strong agreement between the synthetic and measured dataset were observed. The synthetic dataset was able to accurately represent both the feature distributions of the measured parameters and the inter-feature relationships between IOPs and in water constituents. In most comparisons the synthetic dataset reliably reproduced both the central tendency of the density function and its spread. Further evaluation of the synthetic dataset is required to determine the benefit of this augmentation method on retrieval accuracy of deep learning inversion models.

REFERENCES

Aslan, B. and Zech, G. (2005) Statistical energy as a tool for binning-free, multivariate goodness-of-fit tests, two-sample comparison and unfolding, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 537(3), 626–636.

Cavalli, R.M. (2020) Local, Daily, and Total Bio-Optical Models of Coastal Waters of Manfredonia Gulf Applied to Simulated Data of CHRIS, Landsat TM, MIVIS, MODIS, and PRISMA Sensors for Evaluating the Error. Remote Sensing. 12, 1428. https://doi.org/10.3390/rs12091428

Lehmann, M.K., Gurlin, D., Pahlevan, N. et al. (2023) GLORIA – A globally representative hyperspectral in situ dataset for optical sensing of water quality. Scientific Data 10, 100.

Keywords: *Remote sensing, optical water quality, inherent optical properties, diffusion models, data augmentation*