Enhanced prediction of terrestrial feedbacks to the coastal carbon cycle: using machine learning to improve sub-grid biogeochemical processes.

Nick Bouskill[1], Michelle Newcomer[1], Qing Zhu[1], Ben Brown[2],
Kris Bouchard[3], William Riley[1], Eoin Brodie[1],
[1]Climate and Ecosystem Sciences Division, [2]Environmental Genomics and Systems Biology,
[3]Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720

Focal Areas: The paper aims to improve predictability of the coastal carbon cycle through improved model representation and quantification of the terrestrial feedbacks to aquatic ecosystems. The paper covers two focus areas, 1. Improved predictive modeling of terrestrial and aquatic biogeochemistry through ML-enabled surrogate models, and 2. Employing machine learning to integrate multimodal data sets collected across terrestrial and aquatic ecosystems at high spatial and temporal resolution.

Science Challenge: Coastal ecosystems consist of a number of distinct but tightly coupled features, including rivers, estuaries, wetlands and continental shelves. They are regions of disproportionately high biogeochemical activity, and the carbon cycle of the coastal ocean is a major, but dynamic, component of the global carbon budget[1,2]. Extreme hydrological events acting on terrestrial ecosystems play a significant role in regulating coastal carbon-climate feedbacks. Watersheds with high precipitation have higher river discharge of inorganic and organic carbon, nitrogen, and phosphorus. Furthermore, the timing of delivery into aquatic systems is largely dependent on precipitation events in a given watershed. For example, Raymond and Saiers[3] noted that large hydrological events that made up < 5 % of the hydrograph were responsible for nearly 60 % of the annual flux of DOC. This lateral transfer of elements from terrestrial ecosystems largely determines critical coastal characteristics (e.g., the balance between heterotrophic and autotrophic metabolisms), and the trajectory of the coastal carbon cycle. At the present time, the question of whether coastal systems will remain a sink for carbon under future climate or transition to a source of CO_2 remains a critical gap in our knowledge[4,5]. Plugging this gap remains a societal imperative, as climate-change feedbacks to terrestrial ecosystems will result in the increased frequency of drought[6], the onset of wildfires[7], and large nutrients pulse into rivers[8], while continued permafrost thaw delivers centuries old carbon to waterways[9,10]. While several efforts are underway to represent terrestrial-aquatic coupling in Earth System Models[11], the coarse resolution of these models, and lack of mechanistic treatment of carbon and nutrient cycles, and the agents catalyzing these cycles, means that critical nuance in mechanistic fidelity is often overlooked. Furthermore, significant parameter uncertainty of critical ecological processes persists in coarse-scale models.

Exacerbating these issues are the traditional disciplinary boundaries that have resulted in compartmentalized approaches to research, where measurements of carbon and nutrient cycling have either focused on terrestrial processes of retention and release, or downstream impacts on coastal carbon cycling. However, carbon and nutrient flows through ecosystems are unbounded by discipline and need to be researched and modeled as a continuum[12]. Indeed, Tank et al.,[13] recently calculated that of the 5.1 Pg of carbon (C) transferred from terrestrial to aquatic ecosystems globally, less than 1 Pg C reaches the coastal zone, with ~4 Pg being mineralized or buried along the continuum. To this end, new data-driven ML approaches are required that integrate the existing array of sensing datasets (e.g., remote airborne and satellite-based sensing, and local physicochemical sensors) collected across different spatial and temporal scales to more accurately predict how terrestrial and aquatic ecosystems interact.

DOI: https://doi.org/10.2172/1769704

Herein, we outline a framework that, (1) employs ML-enabled, surrogate models to predict rates of sub-grid biogeochemical processes using inputs from coarse scale ESMs, and (2) develop data-driven machine learning packages that account for multimodal datasets including, hyperspectral remote sensing data, and novel automated sensor infrastructure, to develop scale-aware understanding of the seasonality and anomalies associated with carbon and nutrient fluxes to coastal ecosystems, and the feedback to the coastal carbon cycle.

Rationale: Carbon cycling in coastal ecosystems is, to a large extent, underpinned by climate forcing of terrestrial ecosystems. At the present time a number of barriers exist preventing accurate prediction of terrestrial feedbacks on coastal systems. Critical biogeochemical processes occur across different scales, and these scales need to be efficiently bridged. For example, biogeochemical cycling occurs at the pore to aggregate scale (micron to millimeter), where community emergence, as a function of hydrology and climate, determines rates and downstream processes. Furthermore, the microbial engines that catalyze various segments of the carbon, nitrogen, and phosphorus cycle show non-linear behavior in their response to perturbation [14,15]. The ramifications of such activity occur at community to ecosystem level, which eventually determines the watershed function. Accounting for the length and breadth of this continuum, the complexity of responses to disturbance, and the vast differences in scales are critical for accurate prediction of global biogeochemical cycling.

Current land modeling approaches cannot resolve many of the subgrid processes that aggregate towards community biogeochemical cycling. To attempt to do so would exacerbate problems of equifinality, data-storage, and extremely long-run times. Issues that become more complicated as models develop 3D representations of watersheds. Manually tuning land-surface models is also not an option for models of such complexity. However, several studies have demonstrated the utility of ML-enabled surrogate models, for developing simple fine-scale models to inform larger scale models[16], notably to improve atmospheric physics within global circulation models[17,18], and biophysical parameter estimation[19]. The output from these fine-scale models can then be used as parameterization of subgrid scale processes in coarse scale models. While a number of fine-scale models of microbial biogeochemical cycling have been developed[14,20,21], their integration with coarse-scale models is lacking.

Finally, a lack of data explicitly linking terrestrial processes to coastal responses represents a further barrier to constraining our understanding of the terrestrial-feedback on coastal ecosystems. However, data-driven machine learning approaches can be employed to predict the transfer of carbon, nitrogen or phosphorus from terrestrial to coastal ecosystems given information on certain landscape features (e.g., lithography, topography, land cover, precipitation, temperature). The framework would integrate satellite and ground-level data streams to build models relating phenological and climate properties to fluxes of carbon, nitrogen, and phosphorus to coastal ecosystems.

Narrative: *Building surrogate models for terrestrial/ aquatic biogeochemistry:* Surrogate models are necessary to limit run time and domain size while approximating emergent parameters for subsequent coarse model scales. Furthermore, surrogate models reduce the complexity of microbial-centric biogeochemistry modules by learning dynamics directly from observation behavior. Approaches here will reflect those previously devised for Earth System Models[19], however, move beyond simple emulators to build and further utilize sophisticated, fine-scale models of sub-grid microbial biogeochemistry and community emergence (terrestrial bacteria/ fungi/aquatic bacteria/ phytoplankton), run using input from coarse-scale land model variables related to thermal, physical, hydrological properties. To this end we

follow a simple roadmap: 1. *Train*: Build and train a series of deep learning recurrent neural network (RNN) models to predict a series of biogeochemical processes, related to fine-scale carbon, nitrogen and phosphorus cycling networks, using a range of parameter structures and physical constraints (e.g., from no constraints to more detailed functional level constraints). 2. *Emulate:* Use these surrogate models to make predictions of subgrid biogeochemical cycling with increased computational efficiency: here the model inputs would include physico-chemical factors, but also the extent of landscape features of disproportionate biogeochemical significance (e.g., toeslopes or riparian zones) gleaned from the coarse-scale model. 3. *Calibrate:* minimize the error in the prediction of C/N/P fluxes from the terrestrial ecosystem relative to observations (described below) in order to generate best fit parameterization and statistical distributions, and 4. *Testing and uncertainty quantification:* use optimal parameter values and distributions to address whether a more constrained prediction is possible and whether prediction uncertainty could be further minimized by comparison with high-resolution observational data. The relationship generated between coarse scale model input and fine-scale model rates would in turn be used to predict biogeochemical cycling across spatial and temporal scales in order to, (1) improve prediction of retention and release of carbon, nitrogen, and phosphorus within land-models, and (2) improve prediction of the biogeochemical processes within rivers and estuaries that ultimately determine the magnitude of fluxes from the land to coastal systems.

ML-enable data integration approaches: Understanding the feedback from terrestrial to aquatic ecosystems requires improved measurements of the fluxes of carbon, nitrogen and phosphorus from land to waterways. For most watersheds the magnitude and seasonality of these fluxes are unknown. Also unknown is the time lag from the terrestrial response to disturbance to the response in rivers and coastal systems. Constraining these values requires a hybrid approach coupling knowledge-driven, but data poor, understanding of the connections between terrestrial and aquatic ecosystems, and ML-driven data rich approaches[22]. This in turn requires the integration of new data streams, including climate variables, phenological properties (hyperspectral measurements), and soil properties, in a neural network approach to predict carbon, nitrogen, and phosphorus fluxes to coastal ecosystems. Furthermore, deep learning algorithms built upon time-delay neural networks will be used to extract the diurnal, seasonal and annual carbon and nutrient fluxes with and without hydrological extremes. These approaches will improve prediction of carbon and nutrient fluxes out of watersheds by identifying the main causal relationships with the overarching state properties. This approach takes advantage of both remote sensing products for terrestrial and aquatic ecosystems (e.g., Landsat, or Sentinel), and next-generation autonomous sensor packages to monitor the physico-chemical properties in real-time throughout the year. This would extend to two complementary sensor packages: the terrestrial package, measuring soil temperature, moisture, inorganic and organic carbon, nitrogen and phosphorus, and an aquatic package of similar sensor (substituting soil moisture measurements for salinity and turbidity measurements) that can be moored in one place, or allowed to drift with discharge. Finally, it is clear that microbial-centric biogeochemical models suffer from a lack of accurate parameterization, however, there is a wealth of data that can be yielded from whole-genome and metagenomically-assembled genomic data[23]. This data is amenable to deep-learning neural network approaches for extracting traits (e.g., growth rate, temperature optima) that are relevant for predicting community shifts during disturbance, and changes in biogeochemistry.

The products described herein (from ML algorithms to model approaches and data) will be subject to FAIR (Findable, Accessible, Interoperable, Reusable) principles and incorporated into repositories such as ESS-DIVE for public dissemination.

References

1. Laruelle, G. G., Lauerwald, R., Pfeil, B. & Regnier, P. Regionalized global budget of the CO_2 exchange at the air-water interface in continental shelf seas: Continental shelf seas CO_2 fluxes. Glob. Biogeochem. Cycles 28, 1199–1214 (2014).

2. Bauer, J. E. et al. The changing carbon cycle of the coastal ocean. Nature 504, 61–70 (2013).

3. Raymond, P. A. & Saiers, J. E. Event controlled DOC export from forested watersheds. Biogeochemistry 100, 197–209 (2010).

4. Regnier, P. et al. Anthropogenic perturbation of the carbon fluxes from land to ocean. Nat. Geosci. 6, 597–607 (2013).

5. Friedlingstein, P. et al. Uncertainties in CMIP5 Climate Projections due to Carbon Cycle Feedbacks. J. Clim. 27, 511–526 (2014).

6. Brodribb, T. J., Powers, J., Cochard, H. & Choat, B. Hanging by a thread? Forests and drought. Science 368, 261–266 (2020).

7. Higuera, P. E. & Abatzoglou, J. T. Record-setting climate enabled the extraordinary 2020 fire season in the western United States. Glob. Change Biol. gcb.15388 (2020) doi:10.1111/gcb.15388.

8. Rhoades, C. C. et al. The Legacy of a Severe Wildfire on Stream Nitrogen and Carbon in Headwater Catchments. Ecosystems 22, 643–657 (2019).

9. Vonk, J. E. et al. Reviews and syntheses: Effects of permafrost thaw on Arctic aquatic ecosystems. Biogeosciences 12, 7129–7167 (2015).

10. Vonk, J. E. et al. Activation of old carbon by erosion of coastal and subsea permafrost in Arctic Siberia. Nature 489, 137–140 (2012).

11. Lauerwald, R. et al. ORCHILEAK (revision 3875): a new model branch to simulate carbon transfers along the terrestrial–aquatic continuum of the Amazon basin. Geosci. Model Dev. 10, 3821–3859 (2017).

12. Battin, T. J. et al. The boundless carbon cycle. Nat. Geosci. 2, 598–600 (2009).

13. Tank, S. E., Fellman, J. B., Hood, E. & Kritzberg, E. S. Beyond respiration: Controls on lateral carbon fluxes across the terrestrial-aquatic interface: Controls on lateral carbon fluxes. Limnol. Oceanogr. Lett. 3, 76–88 (2018).

14. Georgiou, K., Abramoff, R. Z., Harte, J., Riley, W. J. & Torn, M. S. Microbial community-level regulation explains soil carbon responses to long-term litter manipulations. Nat. Commun. 8, 1223 (2017).

15. Bouskill, N. J. et al. Pre-exposure to drought increases the resistance of tropical forest soil bacterial communities to extended drought. ISME J. 7, 384–394 (2013).

16. Reichstein, M. et al. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195–204 (2019).

17. Rasp, S., Pritchard, M. S. & Gentine, P. Deep learning to represent subgrid processes in climate models. Proc. Natl. Acad. Sci. 115, 9684–9689 (2018).

18. Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G. & Yacalis, G. Could Machine Learning Break the Convection Parameterization Deadlock? Geophys. Res. Lett. 45, 5742–5751 (2018).

19. Dagon, K., Sanderson, B. M., Fisher, R. A. & Lawrence, D. M. A machine learning approach to emulation and biophysical parameter estimation with the Community Land Model, version 5. Adv. Stat. Climatol. Meteorol. Oceanogr. 6, 223–244 (2020).

20. Bouskill, N. J., Tang, J., Riley, W. J. & Brodie, E. L. Trait-Based Representation of Biological Nitrification: Model Development, Testing, and Predicted Community Composition. Front. Microbiol. 3, (2012).

21. Tang, J. & Riley, W. J. Competitor and substrate sizes and diffusion together define enzymatic depolymerization and microbial substrate uptake rates. Soil Biol. Biochem. 139, 107624 (2019).

22. Yang, T. *et al.* Evaluation and machine learning improvement of global hydrological model-based flood simulations. *Environ. Res. Lett.* **14**, 114027 (2019).

23. Vieira-Silva, S. & Rocha, E. P. C. The Systemic Imprint of Growth and Its Uses in Ecological (Meta)Genomics. PLoS Genet. 6, e1000808 (2010).