LIMNOLOGY
and
OCEANOGRAPHY: METHODS

Limnol. Oceanogr.: Methods 19, 2021, 659–672

doi: 10.1002/lom3.10451

How do methodological choices influence estimation of river metabolism?

Anne E. Schechner,[1]* **Walter K. Dodds**,[1] **Flavia Tromboni**,[2] **Sudeep Chandra**,[2] **Alain Maasri**[3,4]
[1]Division of Biology, Kansas State University, Manhattan, Kansas
[2]Global Water Center and Department of Biology, University of Nevada Reno, Reno, Nevada
[3]Department of Ecosystem Research, Leibniz Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany
[4]Department of Entomology, The Academy of Natural Sciences of Drexel University, Philadelphia, Pennsylvania

Abstract

River metabolism modeled from diurnal dissolved oxygen (DO) has become a widely used metric of ecosystem function, yet many papers provide insufficient methodological detail for replication. Only 79% of 43 sampled papers published from 2015 to 2019 mention calibration, 44% describe sensor placement, and 34% did not describe estimation approaches such that the study could be replicated. Given spatial heterogeneity in rivers influences metabolism, and measurement sensitivities vary with sensor model, it is important to have appropriately detailed information in reported methods along with a fundamental understanding of how river heterogeneity might influence metabolism. We deployed 2–8 sensors at 92 steppe river reaches to characterize site heterogeneity, evaluating how sensor placement and type, deployment length, drift correction, data source, local vs. remotely sensed data, and calibration can affect metabolism estimates. Estimates of gross primary production (GPP) and ecosystem respiration (ER) were inconsistent and unpredictable depending on deployment location within a river reach; GPP and ER rates varied up to 131% and 69%, respectively, across a river width and up to two orders of magnitude within a reach. DO sensor brands vary in precision and accuracy; we found even when operated within stated performance range, estimates of GPP and ER could vary by 82% and 198%, respectively, if not calibrated beyond factory setting, as determined using field data from a sample site. Inaccuracies from sensor drift over weeklong deployments led to an average 48% ER overestimation, and 2% GPP overestimation comparing uncorrected with corrected field data. We suggest best practices for more comparable, precise, representative, and accurate methods.

Ecosystem metabolism is central to ecosystem function and is the basis for understanding energy flows and ecological efficiencies from local to global scale. Carbon metabolism consists of carbon fixation (gross primary production [GPP]) and biological carbon oxidation (ecosystem respiration [ER]), as well as their balance (net ecosystem production). These properties describing organic carbon dynamics and rates of activity delineate heterotrophic and autotrophic states in lotic waters (Demars et al. 2011; Dodds and Cole 2007) and are used by managers to assess river condition (Chowanski et al. 2020). Net ecosystem production roughly represents the CO_2 emissions or sequestration by a river where groundwater influence is minimal (Hall et al. 2016). Dissolved oxygen (DO) measurements have commonly been used to measure metabolism in aquatic systems as a proxy for carbon flux; they are less complicated than multispecies bicarbonate equilibria and are detectable against low background concentration, though they do not account for anoxic processes (Dodds and Cole 2007).

We contend many stream metabolism publications do not report sufficient methods to allow replication and confident comparison among studies. We aim to understand how equipment choice, data decisions, and instrument placement in heterogeneous waters influence metabolism determinations. We surveyed methods used in metabolism studies, determined how equipment and data decisions affect resulting rate estimates, and evaluated sensor placement in temperate steppe rivers to understand their influence on metabolic calculations at the local to reach scale. Our primary goal is to quantify potential sources of bias and error from initial experimental design to the final step of reporting methods such that future studies are reproducible and better characterize river metabolism.

Advances in sensors and approaches to estimating metabolism (including iterative Bayesian methods allowing for calculation of error and fit) have allowed broader estimation of

*Correspondence: anne.schechner@gmail.com

Additional Supporting Information may be found in the online version of this article.

stream metabolic characteristics. The more sophisticated models currently applied over the original accounting methods (Odum 1956) are informed by additional metrics including factors influencing aeration and photosynthetically active radiation (PAR) and require deciding between direct measurements or estimation of those metrics. Metabolism estimation also requires measurements of barometric pressure, reach geometry and hydrology, and temperature. Some metrics with strong effects on rate estimates have been considered previously. For example, there is ample literature on estimating aeration, leading some to suggest there are weak relationships among estimates modeled, measured, or calculated from stream and river (hereafter river) morphology and hydrology (Riley and Dodds 2012). GPP has been linked to water velocity (Edwards and Owens 1962), and alongside ER to substrate size and variability (Cardinale et al. 2002). Other metrics might affect GPP and ER estimates but are less completely analyzed. Open access and long-term data sets are increasingly tapped for aggregation and synthesis (Rüegg et al. 2020; Hoellein et al. 2013; Bernhardt et al. 2018) though the quality of those data is not always clear or consistent, but used regardless assuming that the quantity of data outweighs any QA/QC issues with a particular small subset of measured sites.

Many data aggregations do not report key measurement conditions, potentially influencing the reproducibility and interpretation of metabolic rates and factors influencing these rates. For example, StreamPULSE (Koenig et al. 2019), a large, multi-institutional effort that aggregates long-term metabolism data, includes metadata describing reach characteristics, but not the specific location or habitat type where the sensor is placed. Such aggregations generally do not document information on length of deployment, wiper use, and calibration procedures, which can alter precision and accuracy (Hall and Hotchkiss, 2017). We assumed these data aggregations did not document this information because it was not available, which opens the question: how often are key measurement descriptions reported in data sources? Thus, we report a systematic review of the literature as the first step of this paper and explore which characteristics might most strongly influence estimation of rates.

Research questions

This article arose from our experiences attempting to measure metabolism in river segments (reaches) across biomes and continents. We sought to determine if particular river habitats have consistent patterns of GPP and ER and if temporal, logistical, and spatial constraints could influence metabolism. These decisions included equipment choices and methods, temporal and spatial specifics of field deployment, and approaches to data processing. We approached the following questions: (1) Do estimations of rates of GPP and ER vary with data source including sensor type and placement? (2) How does QA/QC influence metabolism estimates? and (3) How can we conduct the most representative, accurate, and replicable metabolism field study? Answering these questions can guide the development of best management practices when quantifying metabolism estimates using DO sensors within rivers. We provide a sample workflow including relevant steps discussed in this article in Fig. 1.

We surveyed the recent metabolism literature to assess most common practices and reporting of key metrics, and to identify potential commonalities and bias. We follow with

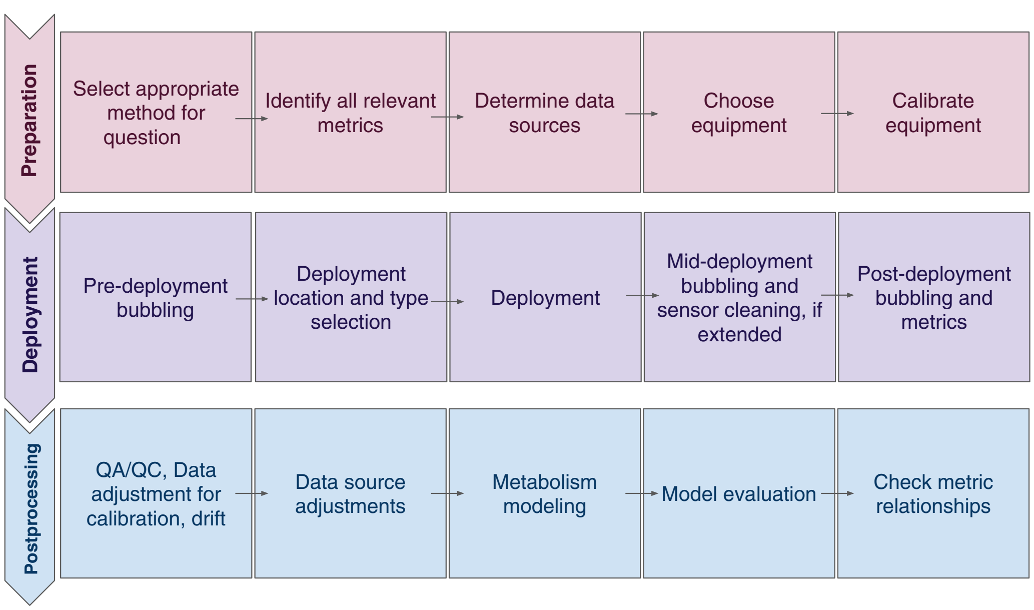

Fig. 1. Sample workflow and decision points of note in open channel one-station metabolism estimation.

analysis of some consequences of inaccurate or imprecise estimation of these metrics. Using field data, we evaluate how estimates can be influenced by sensor precision and accuracy, reliable sampling representative of overall river conditions (e.g., probe placement in the river), and collection in the field vs. remotely obtained values of light and barometric pressure. We demonstrate that individual decision points can change metabolism estimates by orders of magnitude, and that compounded uncertainty, furthered by underreporting, prevents responsible comparison.

Materials and procedures

Methodological review

We examined how recent metabolism papers discussed their methods, and so used a Web of Science (ISI WoS) search on 07 October 2019 with topic parameters [(river OR stream OR aquatic) AND (metabolism OR NEP) AND (diurnal OR oxygen)], refined by [CATEGORIES: (ECOLOGY OR LIMNOLOGY) AND DOCUMENT TYPES: (ARTICLE)], narrowed to years 2015–2019, and focused on the top 140 results. We discarded 96 papers for not using open channel methods, focusing on lakes or fishes, or otherwise not estimating rates of metabolism. We were interested in determining (1) what model or approach authors used to estimate rates of metabolism in rivers, (2) what was the minimum length of sensor deployment, (3) which sensors were used to monitor DO and PAR, and (4) how authors described both calibration and sensor placement. Specific paper titles and information are available in Supporting Information Table S4.

Metabolism monitoring and estimation

We measured DO concentration and temperature at 10-min intervals using Precision Management Engineering miniDOT sensors (Vista, California) for periods between 24 and 144 h, constrained by the logistics of mobile expeditions with multiple stream ecology research objectives in remote locations. At three additional sites, sensors were deployed for up to 2 weeks between calibrations, for a total period of 14 months. MiniDOT files were corrected both for drift and initial calibration based on common average values during pre- and postdeployment logging together in continuously aerated water for at least 30 min to account for any potential change during deployment which could be attributed to biofouling or other causes of instrument drift, including slow adjustment to different conditions including temperature. Our determination of accuracy is based on the assumption that atmospheric oxygen concentrations at each elevation are correct as are the determinations of oxygen saturation concentration dissolved in water as a function of temperature.

PAR was logged at 10-min intervals near the site using Odyssey PAR loggers (Odyssey, Christchurch, New Zealand) calibrated against a LI-COR Quantum Sensor (LI-COR Biosciences, Lincoln, Nebraska) as the manufacturer and Long et al. (2012) recommend. Placement ranged from as close as adjacent to the DO sensor to a few kilometers away at a basecamp, where particularly bare landscapes made locally deployed sensors visible and vulnerable to livestock trampling and/or theft. DO saturation as a function of temperature was calculated using barometric pressure, either measured at the site with a YSI 6020 V2 handheld unit (Yellow Springs Instruments, Yellow Springs, Ohio) or as a daily mean retrieved from the nearest Weather Underground or NOAA station and corrected for elevation. We used multiple YSI 6020 V2 handheld units that were calibrated against each other for barometric pressure and against the NOAA weather station at Manhattan KS Municipal Airport. This allowed us to check if calibration held in the field by comparing multiple calibrated instruments.

Rates of GPP, ER, and aeration were simultaneously estimated alongside standard deviations (SDs) over each 24 h period using the BASE model (v2.3, BAyesian Single-station Estimation, Grace et al. 2015) in R (R Core Team, 2013) run with 200,000 iterations and 10-min interval data, and a θ of 1.07177. We discarded sites where we were unable to model the data with good fit as evaluated by posterior predictive check, modeled vs. estimated data correlation, chain convergence, deviance, and information criteria from the model, as well as a visual evaluation of model fit. We did not estimate the relationship between aeration and discharge in this article as would be necessary for longer deployments in hydrologically variable rivers.

Note the BASE model output is in mass O_2 per volume per time, so results do not rely on accuracy of measurement of river hydrology and morphology (e.g., average velocity, depth, width). If we assume a DO temporal pattern is truly representative of the whole channel, then average depth upstream in the zone influencing the measurement can be used to convert the estimate to per unit area. However, when we place numerous probes in one lateral transect, we cannot know the average depth upstream of the parcel of water above each probe. Thus, our results are reported per unit volume and do not use measures of average depth, which also requires knowledge of average velocity and gas exchange to know how far upstream the measurement was influenced (Demars et al. 2011).

We calculated the upstream zone of influence as the estimated 80% turnover distance as in Hall et al. (2016) (1.61 * Velocity [m d^{-1}]/K [aeration, in d^{-1}]) and report it alongside our metabolism estimates to show how sensitive the calculation of this distance is to the aeration (and velocity) estimation. We did not evaluate or incorporate uncertainty associated with our discharge (and subsequently depth, width, and velocity) sampling methods, but saw no clear increase over our 10 discharge transects. Additionally, we avoided visible lateral or groundwater inflows, which disproportionately affect respiration estimates, as explored in detail by McCutchan et al. (1998). Further, we made multiple discharge measurements along the zone of influence and did not see substantial increases in discharge that would be associated with significant groundwater input.

We used the model put forth by Riley and Dodds (2012) to estimate the initial slope of the photosynthesis-irradiance curve (α) and the maximum rate of photosynthesis (P_{max}) to evaluate differential responses of GPP to light (Jassby and Platt 1975). The relative variation in these metrics is evaluated using the coefficient of variation (CV, the SD divided by the mean).

Hydrology

Velocity profiles were taken at 10 evenly spaced transects over at least a calculated 15-min flow distance upstream of the DO measurement points using either a handheld flowmeter (Marsh McBirney, Hach, Loveland, Colorado) and topset rod at 10 points per transect at 0.6 * depth, or in deeper and non-wadeable rivers using an acoustic Doppler velocimeter (Sontek, Xylem, San Diego, California) pulled across each transect perpendicular to flow direction and corrected to width rather than track distance. River widths were taken at transects, and intermediate points between them, for a total of 19 locations to better characterize variability. We diagrammed site probe placement, indicating characteristics such as relative depth, location along a river width, substrate type, and other relevant details including undercut, bar, and canopy or other vegetation presence.

Study sites

We evaluated river reaches in three ecoregions of the U.S. and Mongolian temperate steppes in summers of 2016–2019, in addition to three locations on the Kansas River for 14 months in 2018–2019. We discuss four of these reaches in detail and include estimates from paired sensors at 23 sites (Supporting Information Table S1). Discharge values ranged from 0.04 to 53 $m^3 s^{-1}$ among sites and captured a wide range of flow conditions. Mongolian rivers had generally open canopies with unstable banks accompanied by heavily grazed riparian zones, and livestock nutrient inputs. Rivers in the U.S. more often had forested riparian zones, flow controlled by upstream impoundments, stabilized channels, established riparian grasses, and cropland nutrient input.

We studied 98 discrete valley-scale hydrogeomorphic units across 19 river networks in temperate steppes of Mongolia and the U.S. These units are delineated as geomorphologically distinct using the GIS-based program RESonate (Williams et al. 2013; Maasri et al. 2019) to extract valley-scale hydrogeomorphic variables from existing geospatial data. We used 10 variables for this delineation extracted at 10 km sample intervals: elevation, mean annual precipitation, valley width, valley floor width (i.e., floodplain), valley width-to-valley floor width ratio, river channel sinuosity, down valley slope, geology, and left and right valley slopes. This approach ensured we had a wide range of river systems for this assessment.

We selected reaches as two riffle-pool-riffle sequences and where in situ hydrology measurements could account for the majority of flow, therefore avoiding braided river sections with more than three parallel channels and river confluence sections. We also avoided reaches in proximity to urban areas, bridges, or other significant anthropogenic features. Reaches had at least a 15 min travel time as calculated by a single velocity transect, with a minimum of 300 m and a maximum of 2 km. We deployed at least two sensors in *an area of active/representative flow*—a phrase we commonly encountered in the literature—but with intent to minimize visibility toward lowering risk of human disturbance. For example, we tied probes to large, submerged rocks or suspended them from overhanging branches. We additionally, based on probe availability at each site, sampled numerous representative or potentially overlooked but contributing "habitats" such as backwaters, debris dams, undercuts, and deep pools. In each of these multiple probe deployments, we used the same type of sensor and calibrated them together before and after deployment to minimize variance not attributable to deployment location. Descriptions of specific placement locations by sampled reach for sites discussed in detail are in Table 1.

We also analyzed a more heavily instrumented site where sensors were placed along a horizontal and vertical transect. We placed four sensors along the surface tied to a wire, and two just above the bottom propped up on rocks, one in the thalweg and the second in an undercut (Fig. 8a).

Assessment

Literature review

We reviewed recent metabolism literature to document the amount of reported methodological detail. Our goal was to determine if the papers followed the basic scientific yardstick of allowing an independent reader to be able to replicate the measurement and subsequent estimation of rates from such measurements. Instrument calibration was mentioned in 79% of papers. Only 44% made any reference to sensor location or attachment point in the reach. About one-third (34%) of papers simply stated that rate estimates were calculated as in Odum (1956); even this classic paper provides seven possible ways of calculating aeration/diffusion. Of the papers that reported using light loggers, 41% stated they used HOBO (Onset Computer Corporation, Bourne, Massachusetts) sensors, while Long et al. (2012) showed HOBO are not cosine corrected (as is standard for estimation of sunlight flux for photosynthetic rates) and have high individual variation. Long et al. (2012) showed that light measures with these sensors could be improved by developing an exponential calibration adjustment from a LiCOR sensor and averaging output from multiple sensors. These papers did not mention using this method, though they may have done so.

Table 1. Characterization of sites discussed where multiple DO probes were deployed to assess spatial effects, as well as metabolism estimates and SDs, and the array reference column links to our figures to denote specific rates and deployment location. $O_{2,80}$ represents the 80% turnover distance as calculated in Hall et al. (2016).

Site	Latitude, longitude	Altitude (m)	Array ref.	Placement description	Discharge ($m^3\ s^{-1}$)	Mean depth (m)	Mean width (m)	Atm. pressure (atm)	GPP ($mg\ O_2\ L^{-1}\ d^{-1}$)	GPP SD ($mg\ O_2\ L^{-1}\ d^{-1}$)	ER ($mg\ O_2\ L^{-1}\ d^{-1}$)	ER SD ($mg\ O_2\ L^{-1}\ d^{-1}$)	K (d^{-1})	K SD (d^{-1})	$O_{2,80}$ (km)
Eg	50.57, 101.53	1121	A	Side, shallow	7.62	0.38	39.8	0.882	1.72	0.08	6.59	0.22	5.97	0.18	11.99
			B	Thalweg					0.91	0.03	3.75	0.08	4.02	0.07	17.78
			C	Side, constrained					0.99	0.02	3.63	0.06	3.58	0.05	19.97
			D	Side, shallow, macrophytes					27.57	8.78	252.75	74.55	26.86	7.89	2.66
			E	Thalweg, constrained, faster					0.85	0.02	3.15	0.07	3.69	0.07	19.38
Tensleep	44.25, −107.22	2709	A	Side, surface	2.57	0.66	10.7	0.732	0.27	0.07	5.83	0.34	5.67	0.28	8.92
			B	Thalweg, surface					0.31	0.09	6.26	0.45	6.07	0.36	8.34
			C	Surface, btw side and thalweg					1.32	0.37	9.23	1.55	8.36	1.27	6.06
			D	Grassy side, surface					0.39	0.10	6.64	0.46	6.16	0.37	8.22
			E	Thalweg, deep					0.92	0.15	4.47	0.42	9.45	0.72	5.36
			F	Deep, undercut					0.72	0.10	7.77	0.46	6.96	0.36	7.28
			G	Thalweg, on log, surface					1.71	0.59	34.03	7.47	32.10	6.87	1.58
			H	Thalweg, deep					0.04	0.05	8.24	1.01	8.87	1.09	5.71
Delgermurun	49.64, 99.92	1284	A	Side, shallow	8.63	0.48	33.72	0.857	1.78	0.03	12.03	0.21	3.02	0.05	19.63
			B	Mid channel, deep					1.74	0.03	10.29	0.16	2.46	0.04	24.06
			C	Thalweg, deep					1.54	0.02	8.65	0.14	2.06	0.04	28.69
			D	Shoreline, slow					1.75	0.04	10.77	0.22	2.63	0.06	22.52
			E	Center, deep					1.67	0.02	9.64	0.15	2.21	0.04	26.86
			F	Side, shallow					1.79	0.03	11.79	0.16	2.60	0.04	22.79
Eg 2	50.1, 101.59	1168	A	Backwater	0.43	0.26	6.68	0.862	1.83	0.13	2.44	0.18	7.14	0.31	6.20
			B	Side, shallow					3.58	0.52	2.37	0.48	20.94	2.15	2.11
			C	Thalweg, shallow					2.06	0.26	11.49	0.94	13.92	1.01	3.18
			D	Thalweg, deep hole					2.21	0.27	15.05	1.20	13.19	0.96	3.36
Tongue	44.77, −107.47	2143	A	Btw side and thalweg	1.78	0.39	13.38	0.732	1.23	0.17	1.13	0.21	11.05	1.04	4.00

Sensitivity analysis: Equipment choice

Equipment choice is an early decision point that could influence quality of estimates; we identified the eight most commonly used sensors in recent literature (based on the literature analysis of 43 papers described in the introduction, citations are provided in Supporting Information Table S4), with accuracies and resolutions as reported by manufacturers (Table 2).

We used a sensitivity analysis to identify how this range of accuracies, up to 5%, might result in different estimates. We found that adding $\pm$ 0.1, 0.2, and 0.4 mg L^{-1} (1.3%, 2.7%, 5.3% difference relative to saturation at the site we used to make this calculation) to each DO reading over 24 h was responsible for a maximum 82% difference in GPP but a 198% difference in ER (Fig. 2) as compared to the calibrated and drift-corrected trace we collected in the field from a relatively metabolically active site. To be explicit, we use percentage difference among separate points to refer to the absolute value of their difference over their average, rather than the equation for percentage change in one point over time, the difference between the final and initial value over the initial value. This illustrates how two probes, both within factory calibration but not calibrated against supersaturated water and/or each other, deployed in the same location could result in substantial differences in estimates of GPP and ER. Our tests assume sensors have similar precision which we therefore assume in our tests, while correcting for drift more directly addresses issues related to accuracy. Repeated measures in the form of multiple sensors should additionally improve precision, but calibration promotes accuracy. We picked and discussed this site on the Tongue river intentionally to illustrate the effects of high aeration and reasonably high GPP. It was, therefore, added in addition to sites discussed in greater detail in this text, and is present in Table 1.

These data make it clear that relying upon factory calibration alone can possibly lead to more inaccurate estimates than those obtained with data generated with careful field calibration procedures, and ER rates may be more strongly influenced by poor calibration than GPP. This probably occurred because GPP is estimated from diurnal changes in DO coupled with departure from saturation, whereas ER is estimated solely by departure from DO saturation.

Spatial sensor placement affected metabolism estimates

The placement of sensors in river channels affected the estimates derived from models for GPP and ER. In particular, one

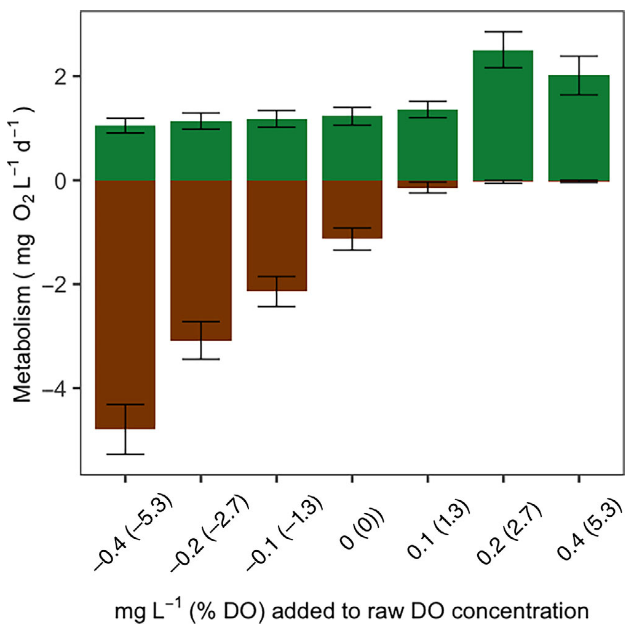

Fig. 2. Rate estimate ($\pm$ SD) variation when each DO reading is adjusted by adding up to 0.4 mg L^{-1} (5.3%), based on one 24-h period on the Tongue River. Specific values available in Supporting Information Table S3.

Table 2. Manufacturer-reported accuracy and resolution of commonly used commercial DO loggers as well as the effects of individual sensor uncertainty on a sample rate from a set of field data used to estimate GPP and ER from one 24-h period on the Zakhvan river. Estimates are derived from adding the amount of DO reported by each company as accuracy to each reading over 24 h, as compared to the estimates of GPP and ER from modeling the actual data based on calibrated, drift corrected probes, both from the BASE model.

Sensor	Reported accuracy between 0 and 8 mg L^{-1}	Resolution	GPP (mg O_2 L^{-1} d^{-1}, as compared to 1.74)	ER (mg O_2 L^{-1} d^{-1}, as compared to 10.29)
Campbell Oxyguard	0.2 mg L^{-1}	0.2 mg L^{-1}	1.77	10.23
Driesen + Kern Logger	0.05%	1%	1.75	10.10
Hach Hydrolab	0.2 mg L^{-1}	0.01 mg L^{-1}	1.77	10.23
Hach Lange	0.1 mg L^{-1}	0.10%	1.85	9.89
HOBO Logger	0.2 mg L^{-1}	0.02 mg L^{-1}	1.77	10.23
PME MiniDOT	5%	0.05 μmol L^{-1}	1.79	10.14
YSI ProODO	0.1 mg L^{-1}	0.01 mg L^{-1}	1.85	9.89
Orion Oxygen Probes	2%	0.1 mg L^{-1}	1.79	10.14

site showed how deployment locations distributed vertically and horizontally across a river cross section can yield widely varying estimates (Fig. 3): over 36 h, four sensors cycled between 8.9 and 9.9 mg L^{-1} daily, while a fifth sensor (Fig. 3, Sensor D) placed in a flowing macrophyte-dominated fine sediment area cycled into hypoxia daily, despite being located within 1 m of another sensor that had a minimum oxygen concentration just below 9 mg L^{-1} (Fig. 3, Sensor A). The other sensors varied substantially from the median calculated while disregarding the macrophyte site, with estimates of GPP differing (Bayesian mean ± SD not overlapping, shown in Table 1) between all but two sensors.

The GPP functional characteristics calculated from diurnal DO curves as in Riley and Dodds (2012) exhibited variability in estimates of α (the slope of the initial response to light) and the maximum photosynthetic rate ($P_{\max}$) for each sensor at two sites (Fig. 4). $P_{\max}$ and α estimates were comparable at all sensor locations at Delgermurun (CV 0.13 and 0.35, respectively) and Tensleep (1.09 and 0.27, respectively). Thus, relying on single-point measures to calculate these photosynthetic parameters may be more reasonable than for metabolism rate estimates.

Absolute and relative ER values varied more than GPP at two of three sample sites (Fig. 5, CV GPP vs. ER, 0.63 vs. 0.24 at Tensleep, 0.05 vs. 0.12 at Delgermurun, and 1.85 vs. 2.06 at Eg). Inconsistent differences appeared when estimates were aggregated by broad characterization as side or center and shallow or deep (Fig. 6). For example, the sensor described in Delgermurun as "Thalweg/Deep" is present in both "center" and "deep" categories.

Sensor placement was important in many sites (Fig. 7). Each end of each line in this figure connects GPP and ER estimates from one probe at a given site with that of a second probe in the same cross section. For example, the line could show a contrast between one shallow and one deep sensor, or one side and one center sensor within the channel. Locations with greater ER were also generally those with greater GPP, but ER rates were greater in magnitude and variability: the median difference in GPP between two paired sensors was 0.72 g O_2 L^{-1} d^{-1} (SD 4.51), while the median difference in ER was 2.23 g O_2 L^{-1} d^{-1} (SD 9.54). Site information for these sensors is detailed in Supporting Information Table S1.

While diurnal DO trends appear similar at a highly instrumented transect (Fig. 8c), rate estimates (Fig. 8b) were sensitive to the apparently minor differences in the diurnal DO (maximum DO varied less than 1% but GPP and ER varied up to 131% and 69%, respectively). It was unexpected that B and C would have such disparate estimates, but we believe

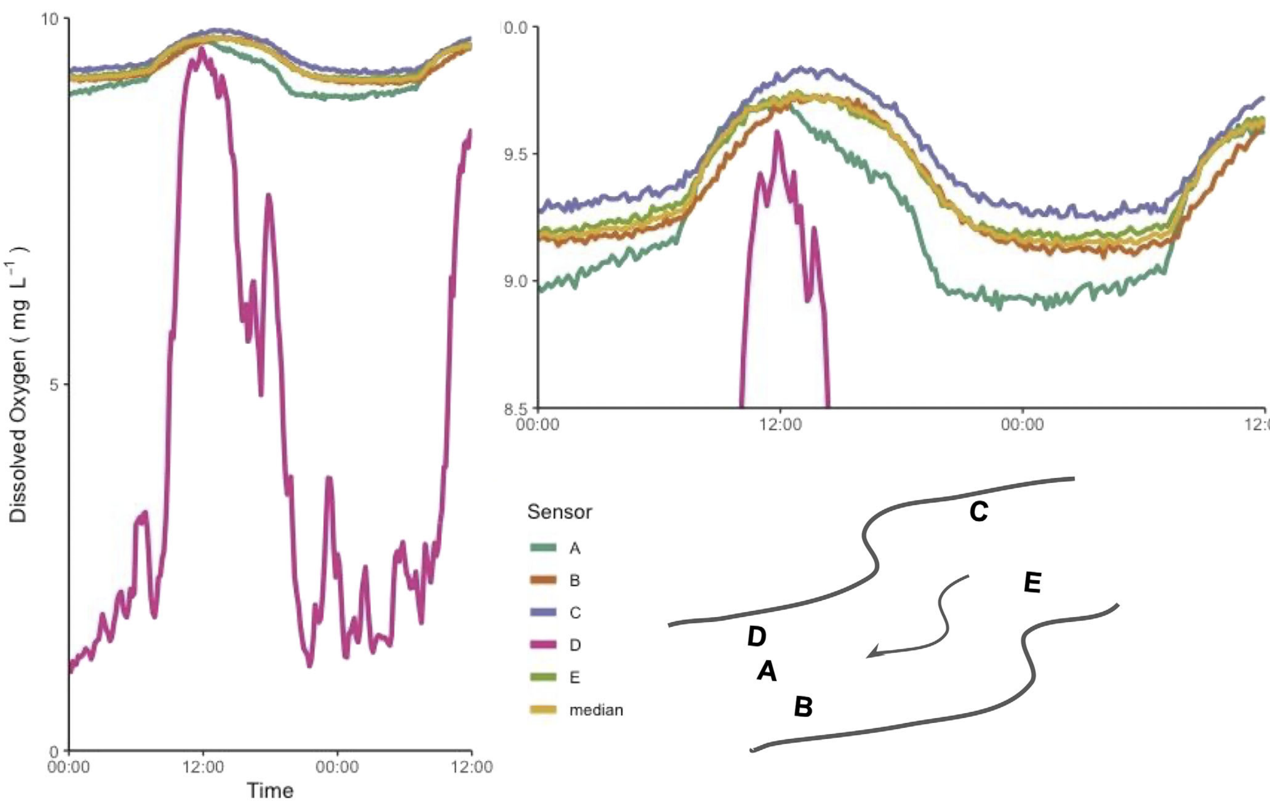

Fig. 3. DO concentrations over 36 h on the Eg river plotted full-scale and zoomed in for five sensors A–E as well as the median of sensors A, B, C, E, and a diagram of their arrangement in the reach. Average reach width was 39.8 m and distance between sensor sets was 575 m.

this difference is due to the fact that B was located in the thalweg and related to C having the biggest error associated with its estimate. We were interested in the fact that sensors B and F also had highest P_{max} despite having the least similar, or rather most exceptional, placements. This, alongside the spread evident in some instances represented in Fig. 7, supports Demars et al.'s (2015) recommendation that multiple sensors be averaged even in well-mixed areas to better incorporate localized heterogeneity in single station estimates.

Refining calibration procedures

We ran all sensors together in DO-saturated water to identify sensors that were clear outliers to minimize bias by equipment as is common in other QA/QC protocols. This

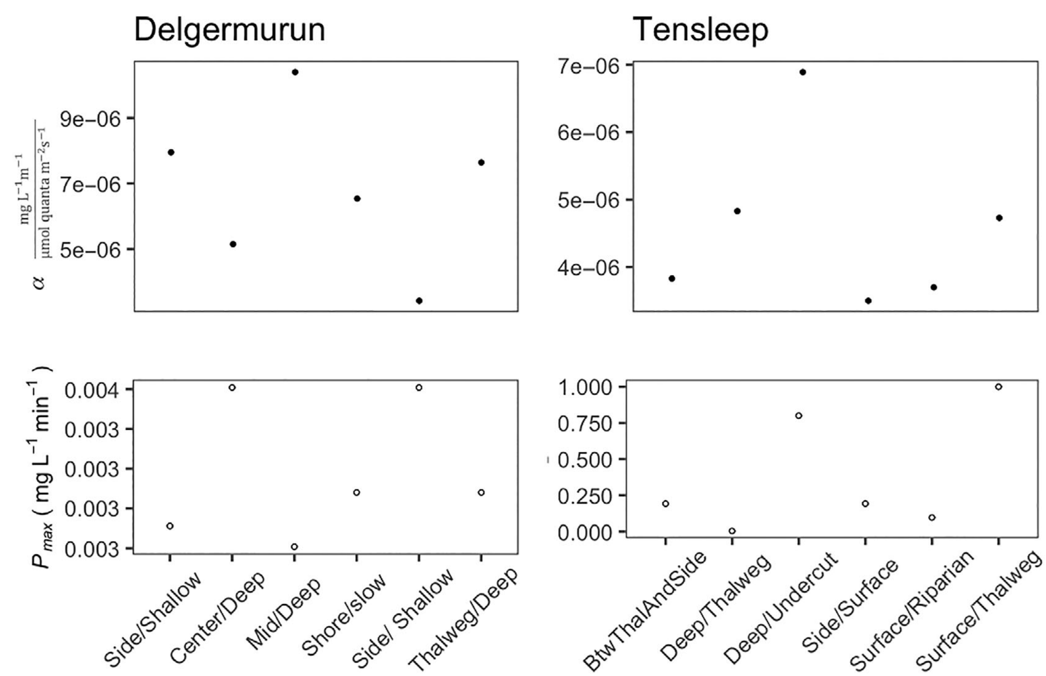

Fig. 4. Rates of α (initial slope of response to light) and P_{max} for each sensor at two sites calculated for one 24-h period. Note differing axis ranges. CV for P_{max} and α for Tensleep were 1.09 and 0.27, respectively, and were 0.13 and 0.35 for Delgermurun.

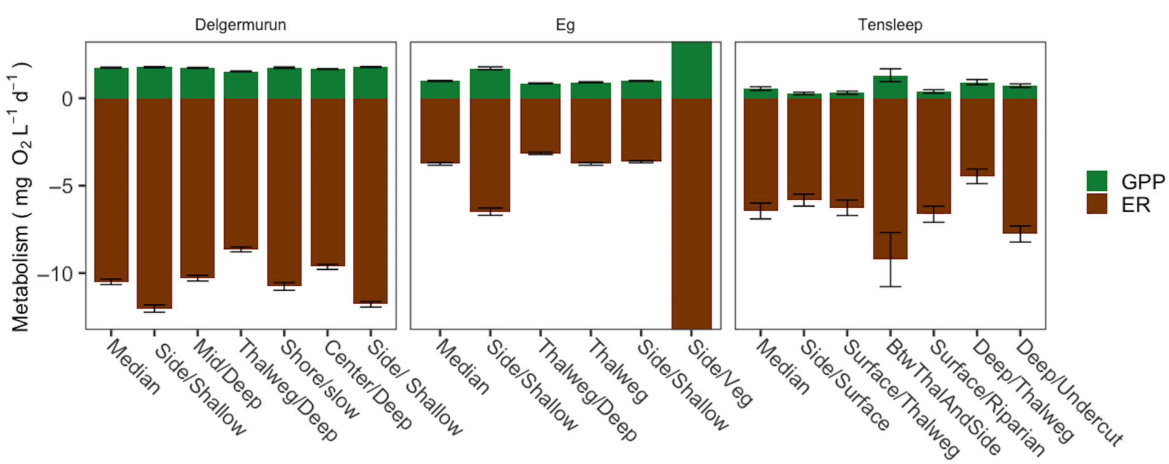

Fig. 5. Volumetric rates of metabolism ($\pm$ SD) at different locations in one area along each river length, as well as their median. The Tensleep site is a direct array (all sensors attached to a cable running across a river width), while the Delgermurun and Eg sensors are at multiple locations within a 15 min travel time reach. The Eg sensor labeled by the "Side/Veg" bar has dramatically greater rates of production and respiration (GPP 27.57 $\pm$ 8.78, ER $-$252.75 $\pm$ 74.55), and is shown in more detail in Fig. 3.

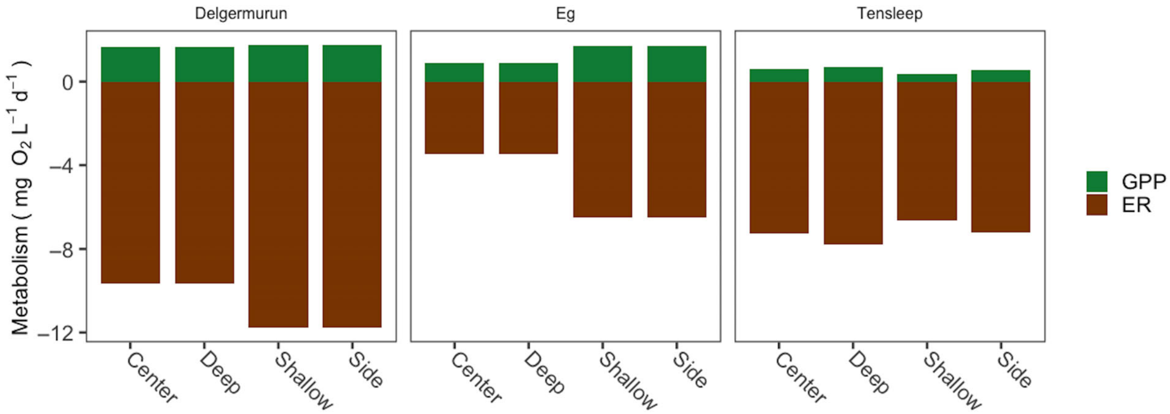

Fig. 6. Median rates by habitat designation as side or center and shallow or deep at three sites. For example, all sites within a river in Fig. 4 that have Side as part of the location description on the x-axis are included in the "Side" median bar.

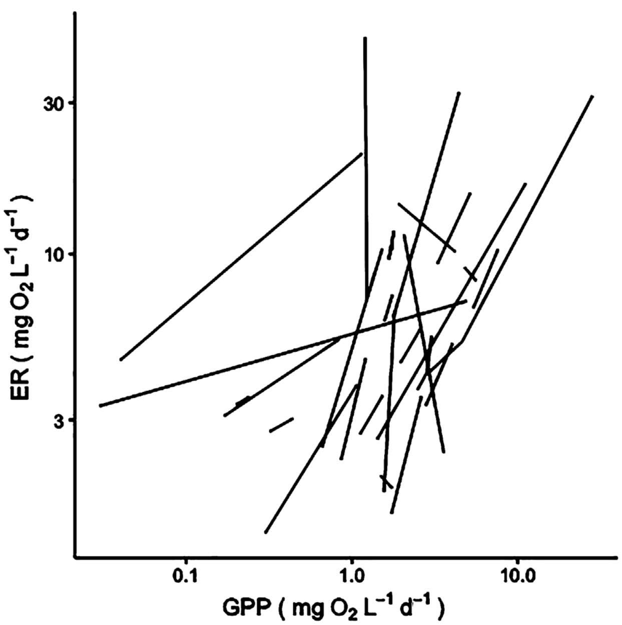

Fig. 7. GPP and ER estimates for each of two paired sensors at a given location (37 pairs at 23 sites, shallow vs. deep or side vs. center, note log scale). Each line connects estimates from two separate probes deployed in the same river transect.

procedure occasionally identified sensors that had varied widely from factory calibration or that were malfunctioning. Such sensors generally could not be calibrated properly. We observed that even with calibration, sensors can drift following deployment (related to, e.g., biofouling or physical changes in the optical dyes used in the sensors over time). Our sensors were not fitted with wipers which could have decreased biofouling, and as such were not suitable for long deployment. We compared estimates for 3 weeks of data from 1 yr at the same site (Fig. 9) with correction based on pre- and postdeployment bubbling to the same data without drift correction. Drift-corrected mean ER varied from uncorrected by up to 100% and GPP by up to 4%. While biofouling may explain drift, we cannot rule out other potential sources, and calibration based on before/after readings corrects for these as well, and should be considered for probes with and without wipers.

Barometric pressure source can affect estimates

Here we assessed differences among potential barometric pressure values. We compared a value collected at the site using a ProODO (Yellow Springs Instruments, Ohio, U.S.A.) handheld unit with barometric pressure sensor, as well as hourly average, daily average, and monthly average from the nearest weather station with historical data as corrected to site altitude (NOAA, Worland, Wyoming, 68 km away). We also used the daily average at the NOAA site and as provided at sea level to evaluate the importance of adjusting to site altitude (2709 m, Fig. 10). As the saturation value of the atmosphere is a function of barometric pressure, not accounting for daily variation of atmospheric pressure could alter results based on changing influx or efflux rates of sub- and supersaturated DO, respectively.

In general, estimates should be based on continuous barometric pressure data, but in a typical day, a single value may suffice if weather patterns do not lead to strong swings in barometric pressure. The variable data estimated lower rates of both GPP and ER compared to the daily mean of the same data. Altitude correction is essential, and even a 500 m difference can have a large impact on measurements.

Light measurement

We note that some of the uncalibrated light probes commonly used do not provide accurate estimates of PAR, though they may have responses to light that are directly correlated with PAR values from cosine-corrected sensors. These data can

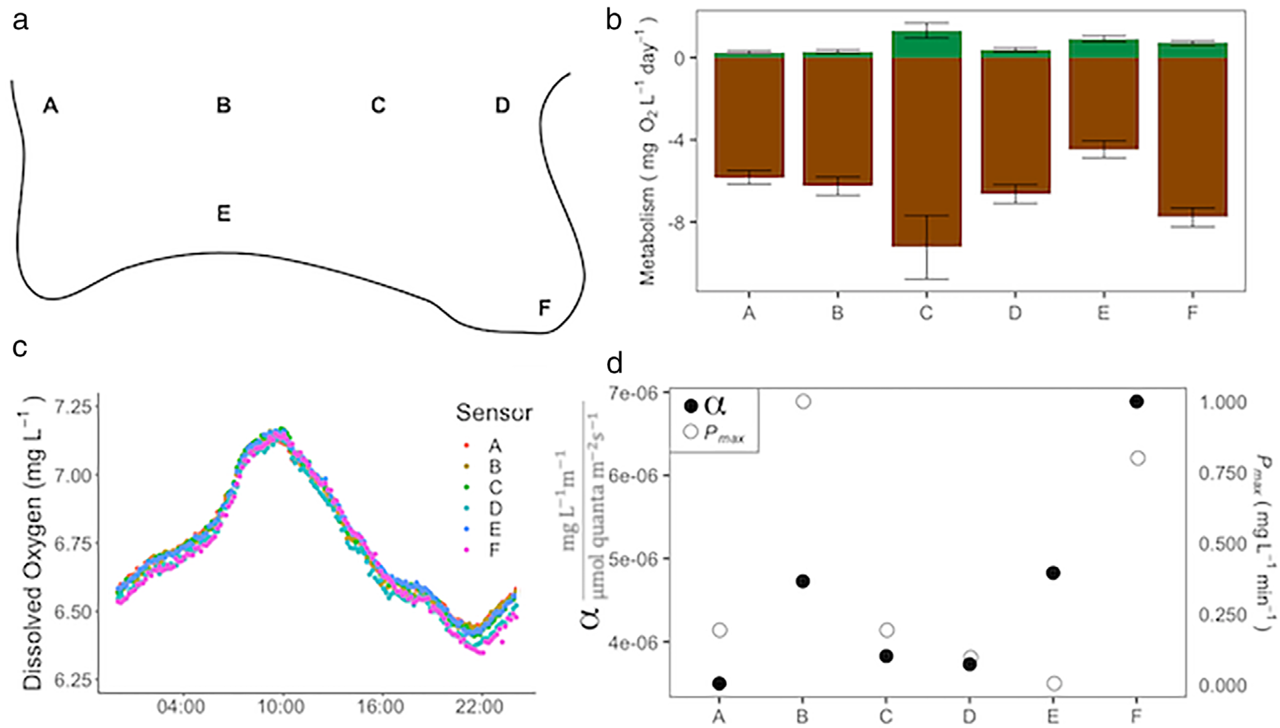

Fig. 8. Cross section sensor array diagram where capital letters represent individual sensors (**a**), 24-h rates of GPP and ER ($\pm$ SD) for each sensor (**b**), overlaid DO concentration over 24 h (**c**), and α , $P_{\max}$ for each sensor (**d**) all at one site and location on the Tensleep River.

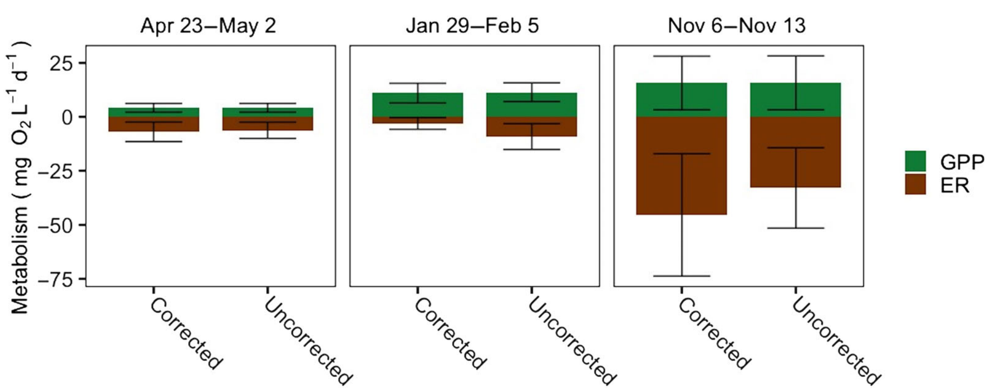

Fig. 9. Average rates of GPP and ER ($\pm$ SD) for three separate full weeks of data in three different seasons of continuous monitoring on the Kansas River near Manhattan, Kansas, with and without drift correction based on pre- and postdeployment runs in oxygen-saturated water.

be used to link diurnal DO traces to GPP by linking changes in light to rates of change in DO. However, if the parameters describing functional relationships of GPP to light (α and $P_{\max}$) are to be investigated and reported such that others can use the estimates, they should be based on calibrated measures of PAR. Bookkeeping models (Odum 1956) would be less affected by this difference as only sunrise and sunset times are relevant, though these times could still deviate from spatially derived values based on local shading. Local shading would not be a problem on an open landscape with modest topographic relief. Models including the BLAM (Julian et al. 2008) can incorporate topographic shading alongside a range of hydrogeomorphic variables to estimate light at the water surface or at depth, but require much additional effort and only averaged accurate within 39% over more than a week of use.

After losing several PAR sensors to theft, we looked to alternatives including placing sensors at nearby and more protected basecamps, as well as by calculating diurnal PAR using geographic location, as included in and recommended by the StreamMetabolizer package (Appling et al. 2017) This model is widely used (Appling et al. 2018; Judd et al. 2009). Modeling light assumes that the location of the sensor is in an unobstructed reach on a clear day, as clouds can dramatically change PAR. When we compared the difference between measured and calculated PAR, we had considerably different estimates for both GPP and ER at one site (Delgermurun 1.74 ± 0.03 vs. 2.10 ± 0.02, -10.29 ± 0.16 vs. -12.64 ± 0.10, respectively), and smaller differences at another site (Eg 0.91 ± 0.03 vs. 0.95 ± 0.02, -3.75 ± 0.08 vs. -3.79 ± 0.07). Total daily irradiance for Delgermurun was measured at 32 mol m^{-2} d^{-1} but calculated by StreamMetabolizer at 52 mol m^{-2} d^{-1}, while at Eg was measured at 53 mol m^{-2} d^{-1} but calculated at 44 mol m^{-2} d^{-1}. Both sites were in relatively flat, open areas, so shading from canopy cover or topography cannot explain this variability. Calculated light curves will miss these interfering factors (e.g., cloud cover, canopy, topography, or other shadows) affecting both sensors and rivers (Fig. 11) that can change estimates. Part of the ability to calculate GPP can be based on DO responses to these shorter-term light fluctuations.

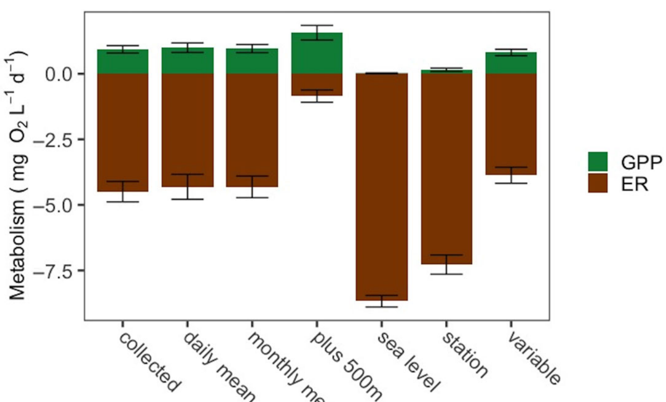

Fig. 10. Estimates of GPP and ER for one 24-h period as vary by barometric pressure source: based on a single handheld sonde measure at the Tensleep site in the Bighorn National Forest, Wyoming (collected), daily mean barometric pressure as obtained from NOAA at the Worland, Wyoming airport 68 km away and corrected to site elevation (2709 m), monthly elevation-corrected mean (NOAA monthly average corrected to site elevation), using hourly variable elevation-corrected data obtained from NOAA, daily mean as would be miscalculated by adding 500 m to the site elevation (plus 500 m), daily mean sea level uncorrected to altitude, and daily mean from the NOAA station in Worland based on station altitude (1239 m). Specific parameters available in Supporting Information Table S2.

Discussion

We show evidence of methodological bias and underreporting in metabolism estimates and literature, but by no means have provided an exhaustive examination of each methodological decision point. We highlight some of the decisions to be considered and attempt to prioritize methodological practices most likely to reflect reality, given site and resource constraints.

We show that system heterogeneity can influence metabolism measures. Heterogeneity has become better appreciated by lotic ecologists as the discipline has matured (Fausch et al. 2002, Frissell et al. 1986), but is still not often directly addressed in whole-river metabolism study, though the heterogeneity on scales from biofilm assemblage to the river continuum has been documented (Cardinale et al. 2002) and may be used (e.g., by incorporating lateral and subchannel inflows) to better assess aquatic-terrestrial linkages and watershed context (Demars 2019). Demars et al.'s (2015) equations to calculate the percentage turnover of DO typically indicate 80% gas turnover rates in hundreds of meters for small streams and in kilometers for rivers. Our data suggest DO measured in well-mixed, main flow areas are most likely to provide results averaging across more upstream heterogeneity, though some sites on the sides or bottoms of main channels can deviate substantially from areas of main flow. Demars (2019) specifically averaged multiple diurnal curves to account for such heterogeneity, noting that this propagates additional uncertainty from each individual sensor.

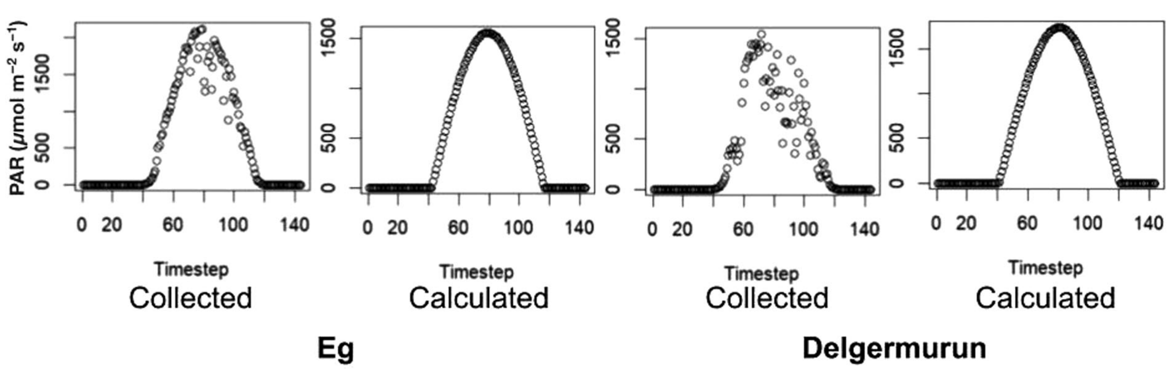

Fig. 11. Collected and calculated PAR at two sites over 24 h, where each timestep represents one reading from each 10-min interval.

Sensor deployment in a location of intermediate depth is also important as thermal stratification can cause the diurnal oxygen cycle to deviate from the main channel and can interfere with obtaining good model fit. This is particularly problematic if the sensor is below the thermocline but also could be important downstream as stratification breaks. Such stratification may be responsible for some of the wide differences in Fig. 7, as sensors were placed in relatively deeper or bankside locations. We found evidence of daily thermal stratification of pools in at least one of our sites, and as such would avoid pools for sensor placement, also discussed in Siders et al. (2017). Additionally, we did not quantify the effects of groundwater inflow (though we saw no visible increases in discharge over the 10 measured transects). Groundwater influx could be part of the reason we saw different oxygen dynamics at a shallow vs. deep sensor at the same thalweg location, consideration of groundwater evaluation as discussed in Hall and Tank (2005) seems prudent.

Including the additional sources of uncertainty we examined in modeling and estimation could be done in the form of additional priors to Bayesian models, or by adding the uncertainty ranges associated with heterogeneity to bookkeeping approaches, improved by Monte Carlo simulations as in Demars (2019). The additional error from any particular metric may seem insignificant relative to the error internalized in our models. We showed sample sites not chosen to represent extreme conditions. Site heterogeneity aside, the sensitivity analysis representing sensor calibration (GPP and ER showing a maximum 82% and 198% difference, respectively), barometric pressure source given correct altitude (20%, 15% difference), light source (19%, 20% difference), and drift correction (4%, 100% difference) demonstrate how this error can compound quickly (summed percent difference 125%, 333%).

The mathematics behind metabolism calculations from DO from a single station measurement assume homogeneity in the channel. In practice, river biogeochemists assume monitoring of 1–2 (usually 1) locations averages all areas and metabolically relevant actors (Hall and Hotchkiss 2017). However, several different scales of heterogeneity may interfere with such averaging. Reichert et al. (2009) and Dodds et al. (2018) document substantial, multiscale metabolic heterogeneity, both examining data from serial reaches and offering empirical approaches to demonstrating heterogeneity and calculating appropriate reach lengths based on upstream influence distance. These data in aggregate suggest careful determination of what constitutes a "representative" reach required to obtain results reflecting general metabolic rates in a river. Few studies we are aware of do this in addition to Reichert et al. (2009), Demars et al. (2015), and Dodds et al. (2018).

Comments and recommendations

We found a number of decision points that influenced metabolism estimates potentially leading to variable estimates. The largest differences in resultant rate estimates in this assessment were associated with deployment location and accounting for sensor accuracy and drift, though α and $P_{\max}$ were more variable within than among sites. Less important were differences associated with saturation calculations (barometric pressure value as long as altitude correction was employed) and categorical channel position (shallow vs. deep, side vs. center designation). ER was more sensitive to most methodological choices than was GPP. McCutchan et al. (1998) found that the greater sensitivity of ER decreased at higher magnitudes of GPP and ER, and reflects larger uncertainty in ER than in GPP in streams with lower rates of each. This result is likely because GPP is driven by diurnal variation and is less affected by aeration uncertainty, while ER estimates are derived directly from the exact difference of DO from saturating concentrations.

We found sensors placed in different areas of active flow gave different rate estimates of metabolism. Sensors placed in the thalweg but off the bottom gave the closest to mean rates. Equipment choice clearly influenced outcomes, improved further by careful calibration and QA/QC procedures. Finally, locally measured vs. remotely sensed light and barometric pressure resulted in different rate estimates, to a lesser degree assuming low topographic relief and accurate altitude.

Calibration of all sensors used for metabolism estimates is important. However, given that many papers do not report calibration protocols, it is difficult to know how to assess the reliability of estimates presented in those papers. Our data show that calibration of DO probes can be one of the most important factors influencing estimates of metabolism.

Increasing reporting of methods increases the value and utility of data. Our review found incomplete methods reporting of sensor preparation, data QA/QC, deployment location, modeling approach, and parameter sources. Any of these differences in methods would have altered metabolism estimates, some substantially. Without including this basic information in papers, data-harvesting initiatives, monitoring networks, and management decision-making, our analysis suggests that the possibility to repeat the measurements is not being met by a considerable portion of the literature. This information is also fundamental for comparative and meta-analysis.

Our recommendations are generally simple enough to adopt, and we empirically show that following general operational guidelines and reporting can improve the value and precise comparison of estimates among studies. In order of priority with respect to measurement methods, (1) Carefully calibrate all sensors before and after deployment (particularly DO sensors), and correct for drift based on calibration before and after the period of measurement. Do not rely on factory "calibrations"; (2) Deploy sensors in or as close as possible to the thalweg. Ensure that sensors are not placed along/are oriented away from the bottom or sides of the channel and not placed in areas with poor mixing with the rest of the channel; (3) Whenever possible use supporting data (light, barometric pressure) taken in the river or as near as possible, and logged at the same frequency as the DO; (4) pay particular attention to the fact that barometric pressure may be reported corrected to sea level; (5) report all measurement and calculation approaches, including calibration procedures, probe placement, data cleaning steps, and programs used for estimation. The percentage error associated with each of these steps may vary with site conditions. We omit many points of consideration that would be crucial for longer-term deployment, which would be best served by a preliminary study evaluating different habitats and upstream zone of influence.

Next steps

Finer spatial scale data could provide a more complete accounting of DO flux across an entire cross section of a river. This estimate could be accomplished by deploying arrays of calibrated DO sensors in tandem with data from an acoustic Doppler velocimeter yielding discharge estimates associated with each DO measurement point. This would allow complete accounting of DO flux for each timepoint, and avoid problems of representativeness. This could also aid in the process of linking the contribution of small habitat differences to reach-scale production and respiration, and in identification of greater or lesser need for increased sampling intensity. This approach would be costly, require substantial effort, and only give an estimate for one cross section of a river system.

A much broader, more detailed, and comprehensive modeling effort may specifically quantify all known error based on synthetic data and a range of possible physical factors including those not discussed here. This effort could be based on observed ranges of aeration, GPP, ER, barometric pressure, temperature, and light variability. This type of sensitivity analysis could more specifically rank the potential biases we document in this current paper, and create response curves of sensitivity of estimations that could be used to refine equipment choices, approaches to sampling and measurement, and data analysis.

References

Appling, A. P., R. O. Hall, C. B. Yackulic, and Arroita, M. Overcoming equifinality: leveraging long time series for stream metabolism estimation. Journal of Geophysical Research: Biogeosciences **123**: 624–45. doi:10.1002/2017JG004140

Appling, A. P., and others. 2018. The metabolic regimes of 356 rivers in the United States. Sci. Data **5**: 180292. doi:10.1038/sdata.2018.292

Bernhardt, E. S., and others. 2018. The metabolic regimes of flowing waters. Limnol. Oceanogr. **63**: S99–S118. doi:10.1002/lno.10726

Cardinale, B. J., M. A. Palmer, C. M. Swan, S. Brooks, and N. L. Poff. 2002. The influence of substrate heterogeneity on biofilm metabolism in a stream ecosystem. Ecology **83**: 412–422. doi:10.1890/0012-9658(2002)083[0412:TIOSHO]2.0.CO;2

Chowanski, K., L. Kunza, G. Hoffman, L. Genzoli, and E. Stickney. 2020. River management alters ecosystem metabolism in a large oligotrophic river. Freshw. Sci. **39**: 534–548. doi:10.1086/710082

Demars, B. O. 2019. Hydrological pulses and burning of dissolved organic carbon by stream respiration. Limnol. Oceanogr. **64**: 406–421. doi:10.1002/lno.11048

Demars, B. O. L., and others. 2011. Temperature and the metabolic balance of streams. Freshw. Biol. **56**: 1106–1121. doi:10.1111/j.1365-2427.2010.02554.x

Demars, B. O. L., J. Thompson, and J. R. Manson. 2015. Stream metabolism and the open diel oxygen method: Principles, practice, and perspectives. Limnol. Oceanogr.: Methods **13**: 356–374. doi:10.1002/lom3.10030

Dodds, W. K., and J. J. Cole. 2007. Expanding the concept of trophic state in aquatic ecosystems: It's not just the autotrophs. Aquat. Sci. **69**: 427–439. doi:10.1007/s00027-007-0922-1

Dodds, W. K., and others. 2018. Spatial heterogeneity and controls of ecosystem metabolism in a Great Plains river network. Hydrobiologia **813**: 85–102. doi:10.1007/s10750-018-3516-0

Edwards, R. W., and M. Owens. 1962. The effects of plants on river conditions IV. The oxygen balance of a chalk stream. J. Ecol. **50**: 207–220. doi:10.2307/2257205

Fausch, K. D., C. E. Torgersen, C. V. Baxter, H. W. Li. 2002. Landscapes to Riverscapes: Bridging the Gap between Research and Conservation of Stream Fishes. BioScience. **52**: 483. doi:10.1641/0006-3568(2002)052[0483:ltrbtg]2.0.co;2

Frissell, C. A., W. J. Liss, C. E. Warren, and M. D. Hurley. 1986. A hierarchical framework for stream habitat classification: Viewing streams in a watershed context. Environ. Manag. **10**: 199–214. doi:10.1007/BF01867358

Grace, M. R., D. P. Giling, S. Hladyz, V. Caron, R. M. Thompson, R. Mac Nally. 2015. Fast processing of diel oxygen curves: Estimating stream metabolism with BASE (BAyesian Single-station Estimation). Limnol. Oceanogr.: Methods **13**: e10011. doi:10.1002/lom3.10011

Hall, R. O., and J. L. Tank. 2005. Correcting whole-stream estimates of metabolism for groundwater input. Limnol. Oceanogr.: Methods **3**: 222–229. doi:10.4319/lom.2005.3.222

Hall, R. O., J. L. Tank, M. A. Baker, E. J. Rosi-Marshall, and E. R. Hotchkiss. 2016. Metabolism, gas exchange, and carbon spiraling in rivers. Ecosystems **19**: 73–86. doi:10.1007/s10021-015-9918-1

Hall, R. O., and E. R. Hotchkiss. 2017. Stream metabolism. *In* Hauer, F. R., Lamberti, G., Methods in stream ecology, v. **2**, 3rd ed. Elsevier. 219–233.

Hoellein, T. J., D. A. Bruesewitz, and D. C. Richardson. 2013. Revisiting Odum (1956): A synthesis of aquatic ecosystem metabolism. Limnol. Oceanogr. **58**: 2089–2100. doi:10.4319/lo.2013.58.6.2089

Jassby, A. D., and T. Platt. 1975. Mathematical formulation of the relationship between photosynthesis and light for phytoplankton. Limnol. Oceanogr. **21**: 540–547. doi:10.4319/lo.1976.21.4.0540

Judd, C., R. Thom, A. Borde, D. Woodruff, C. Roegner, J. Vavrinec, Z. Yang and J. Zhang. 2009. Eelgrass enhancement and restoration in the lower Columbia River estuary. Pacific Northwest National Laboratory, prepared for the US Department of Energy. doi:10.2172/966365

Julian, J. P., M. W. Doyle, and E. H. Stanley. 2008. Empirical modeling of light availability in rivers. J. Geophys. Res. **113**: G03022. doi:10.1029/2007JG000601

Koenig, L. E., A. M. Helton, P. Savoy, E. Bertuzzo, J. B. Heffernan, R. O. Hall Jr., and E. S. Bernhardt. 2019. Emergent productivity regimes of river networks. Limnol. Oceanogr.: Lett. **4**: 173–181. doi:10.1002/lol2.10115

Long, M. H., J. E. Rheuban, P. Berg, and J. C. Zieman. 2012. A comparison and correction of light intensity loggers to photosynthetically active radiation sensors. Limnol. Oceanogr.: Methods **10**: 416–424. doi:10.4319/lom.2012.10.416

Maasri, A., J. H. Thorp, J. Gelhaus, F. Tromboni, S. Chandra, and S. J. Kenner. 2019. Communities associated with the Functional Process Zone scale: A case study of stream macroinvertebrates in endorheic drainages. Sci. Total Environ. **677**: 184–193. doi:10.1016/j.scitotenv.2019.04.394

McCutchan, J. H., W. M. Lewis, and I. J. F. Saunders. 1998. Uncertainty in the estimation of stream metabolism from open-channel oxygen concentrations. J. North Am. Benthol. Soc. **17**: 155–164. doi:10.2307/1467959

Odum, H. T. 1956. Primary production in flowing waters. Limnol. Oceanogr. **1**: 102–117. doi:10.4319/lo.1956.1.2.0102

R Core Team 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Available from http://www.R-project.org/

Reichert, P., U. Uehlinger, and V. Acuña. 2009. Estimating stream metabolism from oxygen concentrations: Effect of spatial heterogeneity. J. Geophys. Res. Biogeosci. **114**: G03016. doi:10.1029/2008jg000917

Riley, A. J., and W. K. Dodds. 2012. The expansion of woody riparian vegetation, and subsequent stream restoration, influences the metabolism of prairie streams. Freshw. Biol. **57**: 1138–1150. doi:10.1111/j.1365-2427.2012.02778.x

Rüegg, J., and others. 2020. Thinking like a consumer: Linking aquatic basal metabolism and consumer dynamics. Limnol. Oceanogr.: Lett. **6**: 1–17. doi:10.1002/lol2.10172

Siders, A. C., D. M. Larson, J. Rüegg, and W. K. Dodds. 2017. Probing whole-stream metabolism: Influence of spatial heterogeneity on rate estimates. Freshw. Biol. **62**: 711–723. doi:10.1111/fwb.12896

Williams, B. S., E. D'Amico, J. H. Kastens, J. H. Thorp, J. E. Flotemersch, and M. C. Thoms. 2013. Automated riverine landscape characterization: GIS-based tools for watershed-scale research, assessment, and management. Environ. Monit. Assess. **185**: 7485–7499. doi:10.1007/s10661-013-3114-6

Acknowledgments

We profusely thank an anonymous reviewer, Scott Ensign, and Benoit Demars who reviewed and much improved an earlier version of this manuscript. We also wish to thank Khaliun Sanchir, Lane Lundeen, Molly Fisher, Samantha Grieger, Andrew Newcombe, Jesús Gomez, Bonnie Teglas, Kari Snelding, Emily Ulrich, and Taylor Hall for field contributions. This research was supported by National Science Foundation grants (NSF Macrosystems 1442562 and 1442544 and NSF EPSCOR OIA-1656006).

Conflict of Interest

None declared.

Submitted 27 April 2021
Revised 20 July 2021
Accepted 23 July 2021

Associate editor: Scott Ensign