# A quantitative & reproducible approach for biophysical connectivity models

Steven Hawes1,2 and Will Figueira1,2  
1 University of Sydney, School of Life & Environmental Sciences (Sydney, NSW, (Australia); 2 Sydney Institute of Marine Science, (Mosman, NSW, (Australia)

## Abstract

*LEAVE THIS TO LAST*

## Introduction

Biophysical models have become a popular tool in advancing our understanding on the connectivity patterns of various marine species {SanvicenteAnorve:2014bl}{Koeck:2015bf}{Andrello:2014ge}. Biophysical models, as the name suggests, produce potential patterns of connectivity using a physical (i.e. oceanographic) model coupled with a biological model. These models are particularly relevant for connectivity in marine systems, as dispersal during a pelagic larval stage is a key driver of connectivity for many species {Cowen:2009fm}. We couple the two models together, because while oceanic processes are the presiding driver of dispersal, biological processes also have highly influential roles {Leis:2006kf}{Pineda:2007bv}. These models are useful as we can predict connectivity patterns over large temporal and spatial scales in a cheap & timely manner, compared to the current empirical techniques.

Utilising to the growing number of connectivity studies using these biophysical models, we wanted to conduct a meta-analysis to answer questions about how early-life history or oceanographic model parameters influence the connectivity and dispersal patterns described in the literature. This was found be an impossible task due to two factors. Comparisons between studies were made difficult due to the variety of metrics used amongst the papers to measure the connectivity and dispersal. The other issue was the lack of transparency on the model parametrisation for many of the studies. The other motivation for the analysis was to validate our own models against existing studies. Validation is an important concern for any modeller and constraints of time and cost make using published literature a sound alternative to empirical research studies.

We reviewed a sample of the literature to see what metrics were used when reporting connectivity studies. The aim was to use these metrics in a meta-analysis, comparing how these metrics changed with parameterisation of the biological model. The second aim was to investigate how parametrisation of biophysical models effected the output. We conducted a database search for connectivity studies, in the last 6 years period (2010-2015) using a biophysical model in their methods. The start date of 2010 was chosen arbitrarily, to limit the search to a manageable amount of papers; note this was not an exhaustive search. The online search engines, ISI Web of Knowledge and SCOPUS, were queried with the keywords connectivity, model\* and marine. The search returned 59 papers fitting within the scope of our study, consisting of 248 authors and published in 31 unique journals. The full list of references is available at (*provide reference to .csv at github perhaps or in supplementary material?*).

## Metrics

### General Introduction

Biophysical connectivity studies measure two distinct – yet related – systems, the patterns of connectivity between natal and settlement sites and the dispersal patterns of the system. The first is essentially a measure of source-sink dynamics, looking at patterns of production and settlement in marine species. The second is a measure of the total dispersal area, looking at issues such as how far the particles travel and where do they occur in highest densities in the system. In undertaking this review, we came across two impediments to a meta-analysis. Firstly, the hardest issue to overcome, was that many studies utilised their own unique metric in measuring either of these patterns. In fact, we found over 30 different measures used to describe the patterns of connectivity & dispersal in these papers. Secondly, often only a single summary statistic was reported, e.g. mean, median, or upper quartile. This is of lesser concern, and can depend on the underlying distribution, but it does make comparisons difficult.

In a field where there can be so many different response variables, it is difficult to nail down a core set of metrics to report and reliably compare against. The most common parameters found in the dataset related to distance the particles travelled and measures of self-recruitment, local-retention and settlement success (Table 1). Although this is not as clear cut as it may seem, as there is confusion about the definitions of self-recruitment and local retention, with these terms often used interchangeably or synonymously with other terms such as self-connectivity. {Burgess:2014cf} stated the definitions clearly and succinctly; local retention is the fraction of offspring produced by a population, which also recruits into that population, and self-recruitment is the fraction of recruitment to a site comprised of individuals born in that site.

The de-facto standards for publishing the results of a biophysical connectivity study in our review, was to use a connectivity matrix as a measure of connectivity between sites (85%) and a dispersal kernel as a measure of spread (71%). A connectivity matrix is a square matrix with rows representing source sites and columns representing settlement sites. The diagonal of the matrix represents self-recruitment and each column, above and below the diagonal shows where settlers are coming from and where particles are settling too. Connectivity matrices are represented as either the potential connectivity (absolute values) or realised connectivity, which is the potential connectivity, normalised by the number of particles released. The most common metrics reported on from these matrices are self-recruitment (70%), local retention (62%), and transport success (42%; Table 1).

A dispersal kernel is simply a probability distribution function of the particles in the system, which gives you a measure of how they moved in the system. Dispersal kernels are often overlaid on GIS images of the study area, in order to give the reader a better understanding of the spread and concentration of the advected particles. The most common metric reported from the dispersal kernel is distance travelled, however, the methods of reporting the distance travelled varies from mean (57.1%), maximum (33.3%), median (11.9%), upper quartile (11.9%), and the minimum distance travelled (4.8%; Table 1). To add to the confusion, there were different interpretations of distance travelled; either it was the distance from the origin site to the final site, or the total length of the particles trajectory. In terms of measuring spread, there was no consensus amongst the studies on how to report the results.

*Do I want to talk about how the different spread metrics?*

*Reference a connectivity matrix & dispersal kernel from another study?*

Both connectivity matrices and dispersal kernels are commonly depicted using heat maps, graphical techniques where a colour system is used to represent numbers. Heat maps are a great visualisation tool for the reader and are to be encouraged, however, it restricts what information which can be extracted from the graph. In general, the range of values depicted are large, and thus they are binned to set of colours, resulting in information loss. Comparative quantitative metrics are essential for analysis between studies, which becomes difficult if the connectivity matrix depicted using a heat map is all that is published as part of the results. Therefore is it imperative for future comparisons that the raw data be published along with the heat maps.

## Models and reproducibility

Reproducibility, a term often synonymous with open science, is an important facet in assessing the scientific merit of a published paper. It can be difficult to achieve in large empirical studies due to experimental variance {OpenScienceCollaboration:2015cn}, however, the same challenges do not exist in computational modelling {Peng:2011et}. Ideally for reproducibility, the model configurations, source code and the output data is made available at the time of publishing. Transparency provides both confidence to the reader, and importantly the peer reviewer, about the outcome and interpretations of the study.

{Sandve:2013gh} produced ten rules of reproducibility for computational research, which all apply to studies involving a biophysical model. Most are intuitive, such as tracking exactly how results were obtained, automating as many steps as possible, and using version control for scripts, code & software. Others are not thought about as often, like recording the seed used for random number generators. The most visible facet to a researcher reading a published study is Rule 10; "all input data, scripts, versions, parameters, and intermediate results should be made publicly and easily accessible" {Sandve:2013gh}.

Transparency is difficult when performing complex modelling because of all the parameters used. Also, researchers are often familiar with either the biological or oceanographic model, making it harder to know what parameters are important to publish. In the subset of connectivity studies that we used for the review, there were many common and easily rectifiable omissions that prevented reproducibility, deduced while inspecting the methods and supporting information. 12% of the studies did not mention the years used for their oceanographic model in their simulation. 34% did not mention the time-step used when advecting the particles in their model. 69% mentioned using diffusion in their model, but only 51% of these published the diffusion coefficient used. When mortality was implemented into the biophysical model, 25% of these studies did not include either the mortality curve or rate used.

Complexity in biophysical models derives from coupling two distinct models, biological and physical, creating an additive effect on the number of input parameters. However, it is impossible to reproduce a biophysical connectivity study without all of these parameters, hence the importance of publishing these parameters along with the results. The strong trend towards supplementary information in journals allows for publication of these parameters without obscuring the primary scientific message. The cost of reproducible research for biophysical studies is minimal, many free online repositories exist for both source code and data, two popular examples being GitHub (http://github.com) and Dryad (http://datadryad.org). Only two studies in the review uploaded their data accessible using such a portal, both choosing to use Dryad. Several journals are also making reproducibility a requirement and providing mechanisms for online storage of data (e.g. PLOS Biology).

## A common approach

In order to provide transparency and allow for useful comparisons between connectivity studies using biophysical models, we need a common approach to publishing both methods and results. We identified two main problems in the literature with connectivity studies preventing us from making meaningful comparisons. The first is the inconsistency in the metrics used in connectivity models and how we present the data. The second is the lack of reproducibility and clarity of the model parameters used in studies.

### Metrics common approach

#### Connectivity matric

Our suggested approach is to utilise the two most common metrics used, these provide useful measures and also they should be readily accepted amongst the research community (Table 2). The first metric, a connectivity matrix, is the most useful for connectivity studies. The measures of self-recruitment, local retention and settlement success can all be calculated easily from the underlying matrix data. Therefore, in addition to presenting the data in a heat map, the full connectivity matrix should be published.

\*ADD VARIABLES USED, can they just attach a connectivity matrix? How about heaps of data?

#### Dispersal Kernel

The other metric is a dispersal kernel. The dispersal kernel gives up valuable data on the spread of the particles and the distance they travel, including the density.

dispersal variables: mean, median, max what are metric should we use for spread??

### Reproducibility common approach

We have supplied a non-exhaustive list of the sort of parameters required to reproduce these studies (Table 3).

## Summary

The main take home message is connectivity modelling studies need to start publishing quantitative metrics of both dispersal and connectivity; too often studies have just focused on one of these facets. We hope, in addition to providing these measures, researchers will also feel comfortable about publishing the data behind their connectivity matrices and dispersal kernels, allowing for future meta-analyses. Research grants are increasingly asking for publication of datasets and as such, open data publications should start becoming standard practice in the near future.

## Acknowledgements

???

### Tables

Table 1: Those most common metrics published in the connectivity studies and the parameters reported per metric used.

|  |  |  |
| --- | --- | --- |
| Metric | Measure parameters | Reported(%) |
| Connectivity matrix |  | **84.6** |
|  | Self-recruitment | 70.0 |
|  | Local retention | 62.0 |
|  | Settlement success | 42.0 |
| Dispersal kernel |  | **71.2** |
|  | Mean distance travelled | 57.1 |
|  | Maximum distance travelled | 33.3 |
|  | Median distance travelled | 11.9 |
|  | Upper quartile distance travelled | 11.9 |
|  | Minimum distance travelled | 4.8 |
| Graph theory |  | **10.2** |

Table 2: The five measures that are required to make comparative analysis from studies using biophysical connectivity models.

|  |  |
| --- | --- |
| Metric | Measure parameters |
| Connectivity matrix | Self-recruitment |
|  | Local retention |
|  | Settlement success |
| Dispersal kernel | Spread |
|  | Mean distance travelled |
|  | Median distance travelled |
|  | Maximum distance travelled |

Table 3: List of input parameters to a biophysical model that should be published as part of every connectivity study using a biophysical model. All values should be published in the appropriate SI units where applicable.

|  |  |  |
| --- | --- | --- |
| Model | Name | Description |
| Particle | Particle model | Link to the source code of the particle model used and specifications with information on type, movement and integration methods e.g. Individual based model, with Lagrangian model and it uses Runge-Kutta fourth order method for integration |
|  | Random number seed | If random numbers are used, it is important to publish the seed used to generate the random number so the study can be replicated {Sandve:2013gh} |
| Physical | Oceanographic model | Link to the specifications of the model, including all sub-models, bathymetry & tidal models used |
|  | Spatial scale | The area the particles can move |
|  | Time step | The time step used to move the particles using the oceanographic model (preferably in seconds) |
|  | Diffusion / turbulence | Diffusion strategy (random walk, horizontal and/or vertical) and the turbulence values used |
| Biological | Release times | Release times of the particles |
|  | Release strategy | How the particles were released (e.g. daily, lunar, temperature cues, depths) |
|  | Release sites | The GPS locations and sizes of the release sites |
|  | Settlement sites | The GPS locations and sizes of the settlement sites (if different from the release sites) |
|  | Particles released | The number of particles released |
|  | Pelagic larval duration | The pelagic larval duration of the taxa and whether you used a variable or fixed duration |
|  | Pre-competency period | The pre-competency period given before settlement could occur |
|  | Movement strategies | What movement strategies were the particles given and the values associated with these strategies (e.g. passive, diel, horizontal) |
|  | Mortality strategies | The mortality strategy used (if any) |
|  | Settlement strategies | How did the particles choose a settlement site (e.g. passively, sites were buffered 10 kms, sensory cues) |

## References