# A quantitative & reproducible approach for biophysical connectivity models

Steven Hawes\*1,2 and Will Figueira1,2  
1 School of Life & Environmental Sciences, The University of Sydney, Sydney, NSW 2006, Australia; 2 Sydney Institute of Marine Science, Mosman, NSW 2088, Australia

Corresponding author: tel: +61 439 881 324; fax: +61 2 9351 6713; e-mail: steven.hawes@sydney.edu.au

Keywords: connectivity, biophysical models, reproducibility, dispersal kernels, connectivity matrix

## Abstract

Biophysical models are an increasingly popular tool in studying marine population connectivity. These models offer the ability to perform analyses of connectivity patterns over temporal and spatial scales that are not possible with empirical techniques. In addition, biophysical models allow for the investigation of the early-life history traits of marine species that influence connectivity. The proliferation of biophysical transport modelling studies should allow for analysis of emergent generalities about the fundamental parameters and processes driving population connectivity which can then served to inform conservation and management initiatives. In a review of 78 recent studies (CITE review; published between 2010-2016) we found two key issues to be limiting our ability to synthesise results from across this diverse array of studies; 1) there was a general inconsistency in reporting of response metrics from these models; and 2) there was a lack of transparency in the values of parameters used as input to these models, creating a lack of reproducibility. Here we provide a summary of the pitfalls created by these two issues and importantly suggest a common approach to the metrics that are reported and a list of parameters to biophysical connectivity studies that are essential for reproducibility. We feel having a consistent approach will allow us to mature the field and encourage greater confidence in connectivity studies. This is important as connectivity studies become drivers of management decisions such as designing marine protected areas or the conservation of specific sub-populations within a range.

## Introduction

Biophysical models (also known as coupled physical-biological models) have become a popular tool in advancing our understanding of the connectivity patterns of marine species with a pelagic larval stage {SanvicenteAnorve:2014bl}{Koeck:2015bf}{Andrello:2014ge}. These models are not a new technique, first appearing in the literature almost 30 years ago {Bartsch:1988vg}. It is an area of research that is becoming more widespread as high-powered computing resources become cheaper. A review of the 17 year period between 1989-2006 found 69 studies that utilised biophysical models {Miller:2007iu} and in our own seven year search from 2010-2016, we found 78 studies fitting this criteria (site our review paper?). Biophysical models 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}. The two models are coupled together, because while oceanic processes are the presiding driver of dispersal, biological processes also have highly influential roles {Leis:2006kf}{Pineda:2007bv}.

Biophysical models are powerful research tools, allowing us to investigate potential connectivity patterns over large spatial scales {Andrello:2013cc}{Treml:2015bt} and over long temporal scales {Richar:2014cg}{Berglund:2012cr}. These are scales that empirical studies cannot capture. These models also provide us the ability to investigate how early-life history traits (e.g. mortality, swimming ability, and orientation) affect the patterns found in connectivity studies {Butler:2011wu}{Staaterman:2012ek}. It allows us to ask important question about how these factors influence connectivity; experiments that are largely intractable *in situ* due to the inherent difficulties of studying the larval phase of marine species.

This 30 year history of connectivity research using biophysical transport models is filled with numerous examples of major findings, advances and generally high quality studies. As with any cohesive field of research, there is a collective aim to advance our overall understanding of the discipline. For instance, what are typical levels of connectivity that one can expect for different types of system and life histories? Are there generalities to the importance of factors like larval behaviour (swimming ability, sensory ability, cues that initiate settlement) or even aspects of the oceanographic forcing models (time steps, resolution, integration method) that profoundly affect connectivity outcomes. Does a longer pelagic larval duration (PLD) increase the size of the dispersal kernel or can a species limit how far they disperse through vertical migration? The establishment, where possibly, of generalities to these and other issues allows the field to progress. It provides guidance for new studies, allowing them to build on previous ones. This process is critical to the use of biophysical connectivity model outputs in management where there is an expectation that they will be accurate and useful across a broad range of scenarios.

While the long history of excellent research in this field provides the substance for some excellent synthetic understanding, we would suggest that our progress in connectivity modelling is being restrained by limitations in two key areas of published studies: 1) inconsistencies in the reporting of response metrics and 2) lack of model and parameter transparency. Of course this is ignoring other limitations of these models, such as the validity of the model output, which we have dealt with in our review paper (CITE Review). In order to evaluate this supposition we sampled biophysical connectivity studies published between 2010 and 2016. We chose to evaluate this recent period as it would represent the ‘best case’ scenario in terms of data availability and of course the most realistic case of how the various aspects of these studies will be reported in the near future. Recent studies also utilise the latest oceanographic models, implementing high resolution models which can capture small scale oceanographic processes.

For each study in the review we recorded its location and scope, the input parameters for the physical model (e.g. the hydrodynamic model used) and biological models (e.g behaviours implemented, spawning times), the response metrics and analyses used. There were several different aims for these connectivity studies; modelling dispersal & settlement of marine species (77% of studies), assessing the connectivity of marine protected areas (15% of studies), investigating how early-life history parameters shape dispersal (5% of studies), and investigating the effects of climate change on connectivity (3% of studies). Most studies modelled a specific species (covering a wide variety of taxa), only five studies took a generic approach to input parameters. Fish were the most common taxon modelled (44% of studies), followed by molluscs (24%) and crustaceans (10%). Below we discuss our findings from the reviewed studies for each of the two key issues we highlighted above.

(Should I look at the categories that Miller 2007 used? Explanatory, Inferential, Hypothesis-generating)

## Metric consistency and reporting

A trait common to nearly all biophysical connectivity studies is that they are attempting to estimate at least one of two fundamental and general patterns; 1) levels of connectivity between natal and settlement sites and 2) dispersal of particles in the system. The first is essentially a measure of larval source-sink dynamics, looking at patterns of production and settlement in marine species with a pelagic larval phase that inherently links one patch in the system to another. The second is a measure of the total dispersal area, looking at issues such as how far the particles travel and where they occur in highest densities in the system. Both traits are useful in describing system connectivity and dispersal and are a product of the system and modelling environment. In undertaking a review (Reference the review), we came across two impediments to a meta-analysis of these estimated metrics. Firstly, 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 78 papers. Secondly, for a given metric, often only a single summary statistic was reported, e.g. mean, median, or upper quartile. Summary statistic are useful, however they were often reported without information on the underlying dispersal distribution, making rigorous qualitative comparisons difficult.

The most common metrics of connectivity and dispersal found in the dataset related to the distances travelled by the larvae particles (dispersal metric) and connectivity measures of self-recruitment, local-retention and settlement success (@tbl:common). This commonality was useful and is no-doubt a response to the focus on local retention of larvae in coral reef literature over the past 15 years {Jones:2005jj}{Berumen:2012ex}{Buston:2012jq}. But even here there was 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.

(Thinking I should put concrete definitions into a supp table?)

Those most common metrics published in the connectivity studies evaluated in this review and the parameters reported per metric used. {#tbl:common}

|  |  |  |
| --- | --- | --- |
| Metric | Measure parameters | Reported(%) |
| Dispersal kernel |  | **64.0** |
|  | 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 |
| Connectivity matrix |  | **57.2** |
|  | Self-recruitment | 70.0 |
|  | Local retention | 62.0 |
|  | Dispersal success | 42.0 |
| Graph theory metrics |  | **8.1** |

The three most common ways of reporting connectivity were metrics to do with dispersal kernels (64.0%), connectivity matrices (59.3%) or graph theory (8.1%). A dispersal kernel is a density plot of the distance travelled by individual particles, providing a measure of the distance dispersed in the system. Dispersal kernels are often overlaid on GIS images of the study area with heat map depicting the density distribution, 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 summary statistics used vary from mean (57.1%), maximum (33.3%), median (11.9%), upper quartile (11.9%), and the minimum distance travelled (4.8%; @tbl:common). These usefulness of a given metric will change depending on the distribution of the dispersal kernel, which is shaped by both geographic features and prevailing currents. Thus in most cases, the omission of any mention or visualisation of the general shape of the distribution (65% of studies that included a dispersal kernel) made it quite difficult to interpret the utility of the reported metric describing it. For instance average distances travelled may be of little use where the underlying distribution is quite right skewed (common in dispersal studies). There was also no consensus on how best to report the breadth of the dispersal distance distribution with studies reporting various metrics such as a dispersion index (defined as the number of particles in an area, multiplied by the distance from the natal site and normalised by the particles released; {Pearce:2011fa}, positive area (defined as the sum of the particles in a unit area; {Andrello:2014ge}, and isotropy (defined as measuring the variance of distribution around the mean geographic location of the particles; {Ayata:2010hg}.

The next most common metric for publishing the results of a biophysical connectivity study in our review, was to use a connectivity matrix (also known as a transition matrix) as a measure of connectivity between sites (59% of studies) and a dispersal kernel as a measure of spread (71% of studies). A connectivity matrix is a square matrix with rows representing source sites and columns representing settlement sites. The diagonal of the matrix represents local-retention 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 (percentage), which is the potential connectivity, normalised by the number of particles released. Self-recruitment is the diagonal of the matrix divided by the sum of the row of the matrix. If egg production varies amongst sites, then the values need to be weighted for comparison. The most common metrics reported from these matrices are self-recruitment (70% of studies with a connectivity matrix), local retention (62%), and dispersal success (defined as the percentage of particles that successfully settle; 42%; @tbl:common).

Another approach is to use graph theory to describe the connectivity (10.2% of studies). Treating sites as nodes, powerful metrics can be calculated to give a measure of connectivity between sites. Graph theory has the advantage of being a mature field in mathematics and it is a useful tool. It is somewhat surprising that a graph theory approach is only introduced in a small number of the studies. Graph theory provides a wide suite of metrics to measure the connectivity between sites, providing the ability to identify connected clusters within the modelled site, the location of the centre of the network, and also the strength of the connections between site. However, like connectivity matrices, it requires the underlying data from the network graph for us to be able to get a more nuanced understanding.

Both connectivity matrices and dispersal kernels are commonly depicted using heat maps, graphical techniques where a colour system is used to represent numbers. For connectivity matrices heat maps can easily identifying larval source and sink sites in the system,and for dispersal kernels they are useful in depicting where the highest and lowest densities of dispersal occur. 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 for further meta-analysis. In general, the range of values depicted are large, and thus they are binned for visualisation, resulting in information loss. As an example of the difficulty in known the underlying values, consider the heat map (@fig:heatmap). Of the studies that published a data from a connectivity matrix, only 35% supplied the actual matrix data as represented in @fig:heatmap. 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.

The probability of larval transport and settlement (as a % of total larvae released from a site) amongst 10 sites (S1-S10) within a simple metapopulation (a) and a simplified heat map of this same data (b) which would be typically used to present the data in a publication. Note that connectivity values derived from the heat map could be up to 50% different from actual values under the binning scheme used here.

The probability of larval transport and settlement (as a % of total larvae released from a site) amongst 10 sites (S1-S10) within a simple metapopulation (a) and a simplified heat map of this same data (b) which would be typically used to present the data in a publication. Note that connectivity values derived from the heat map could be up to 50% different from actual values under the binning scheme used here.

(It was argued this was a bit simplistic and perhaps not required?)

## 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, which was not seen amongst the connectivity studies. 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}. Ignoring the reporting of the random number seed and providing access to source code, only 34% of the studies in our review could be considered reproducible.

Transparency is difficult when performing complex modelling because of all the parameters used. In our review, there were many common and rectifiable omissions that prevented reproducibility, deduced while inspecting the methods and the supporting information (if available). 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 either did not include the mortality curve or rate used. It is advised, where possible, to use oceanographic models that are freely available to the public. Fortunately, this is happening more, whether it is government funded ocean models becoming available through legislative requirements or studies using open hydrodynamics models such as the Regional Ocean Modeling System (ROMS) or the Princeton Ocean Model (POM), which can be coupled with higher resolution models for specific site locations. Transparency amongst models (providing source code) allows for other researchers without the necessary coding skills to utilise the system. There are several examples of this already (e.g. Icthyop, CMS, MGET), suggesting popular support for model reuse amongst the community.

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 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 a lack of commonality of metrics used to describe the connectivity amongst studies. The second issue is the lack of reproducibility and clarity of the model parameters used in studies.

To get consistency amongst metrics, our suggested approach is to utilise the two most common metrics currently reported. These provide useful measures and also they should be readily accepted amongst the research community (@tbl:metrics). The first metric, a connectivity matrix (also referred to as a transition matrix), is the most useful for connectivity studies. The connectivity measures of self-recruitment, local retention and settlement success can all be calculated from the underlying matrix data (means of these values are also useful in publication for easy reference). Therefore, in addition to presenting the data in a heat map visualisation, the full connectivity matrix should be published. It is suggested these are published in an online data-repository or attached as supplementary information, rather than in the body of the paper. This is especially important where studies may have simulated a number of connectivity matrices for different situations (e.g. time periods, parameter combinations). The supplied matrices can be the summary of your data, e.g. if your study was conducted over a five year period, the matrix can be the mean connectivity over this period. It is understood reporting of connectivity matrices can become non-trivial depending on the number of sites used in the model and the number of days larvae are spawned. There is no one solution, instead options ranging from complete data dumps, to summarising results into more management segments (e.g. years), or even only providing data relating to the important findings of the paper. However, if you vary a parameter such as PLD, different matrices should be supplied for each value used value used in order for such results to be more broadly useful to the field in the future. Graph theory is also a useful technique to assess the connectivity of a system. There are easily derived and comparable metrics, however, it can be hard to calculate the underlying values of the connectivity matrix, which makes it hard to compare against studies that do not also use a graph theory approach. So even where graph theory has been utilised, publishing the basic connectivity matrix is still important.

The two metrics that are required to make comparative analysis from studies using biophysical connectivity models. From each of these metrics, four important measures can be derived for further analysis {#tbl:metrics}

|  |  |
| --- | --- |
| Published metric | Derivable measure parameters |
| Connectivity matrix | Site-site connectivity values |
|  | Self-recruitment |
|  | Local retention |
|  | Dispersal success |
| Dispersal kernel | Mean distance travelled |
|  | Median distance travelled |
|  | Maximum distance travelled |
|  | Measure of spread |

The second most important metric used to describe the output of these models is the dispersal kernel. The dispersal kernel——a probability density function——gives information about the spread and distance of dispersed the particles in the system. For the dispersal kernel, a publishing the density plot, along with a heat map, is the recommendation. From a density plot we can derive the mean, median and maximum distance travelled to allow for comparisons between studies, along with the distribution of the dispersal kernel. Note that when we talk about distance, we are referring to actual distance travelled, therefore when other distances such as straight-line or alongshore distances are used it is important to define this. There is no consensus for a measure used for spread. A few of studies in our review used a measure called positive or seeded area, which counted all the cells in the system that had greater than one particle to get a measure of spread. Another two studies tried to calculate the centre of the mass of particles to see how it changed over time. The positive area measure holds potential though it would need to be standardised to the size of the cells used in the model.

To provide transparency between studies on how the biophysical model is implemented, we have provided a non-exhaustive list of the sort of parameters required to reproduce these studies (@tbl:open). The list covers parameters used for both the physical and biological models. This is not definitive, models will increasingly become more complex and there will be more parameters added to the initialisation of the models. For reproducibility purposes, an individual set of parameter values should be supplied for every model run used in the study. Again, these can be supplied in the supporting information or online in a data repository.

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. The supplied column list the percentage of studies that included this information at the time of publication {#tbl:open}

|  |  |  |
| --- | --- | --- |
| 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) |

## Conclusion

In order to advance our understanding of connectivity in marine systems, biophysical modelling studies need to start publishing consistent quantitative metrics of dispersal and connectivity so that comparisons can be made. Studies must also provide more transparency on the parameters used. We have attempted to provide a template for the basic information we think necessary based on our review. We suggest this template be used as the basis for discussion amongst scientists in the field, but strongly encourage all researches publishing the results of biophysical connectivity studies to adopt at least the principles of cohesion, consistency and transparency which are at its core. We feel that such an approach will greatly speed advancements in our understanding of the drivers and outcomes of connectivity more generally. Ultimately it is confidence in the generality of the results of these models which, along with more empirical validation of the modelled output, will allow for their more widespread use within management and conservation initiatives.

## Supplementary materials

The complete list of references in the review are available in Supplement1.docx I might add a glossary to the supplementary material, just to make the definitions clear.

## References