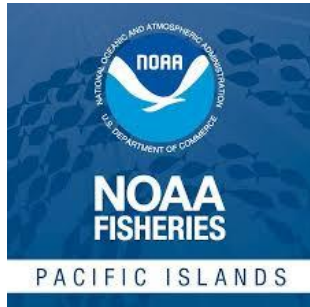


The Value of Diagnostics in Stock Assessment

Felipe Carvalho

Henning Winker

NOAA
Pacific Islands Fisheries Science Center



Joint Research Centre (JRC)
European Commission



Outline

- Motivation, models and data
- Some current diagnostics tests
- Diagnostics innovation, adoption and diffusion

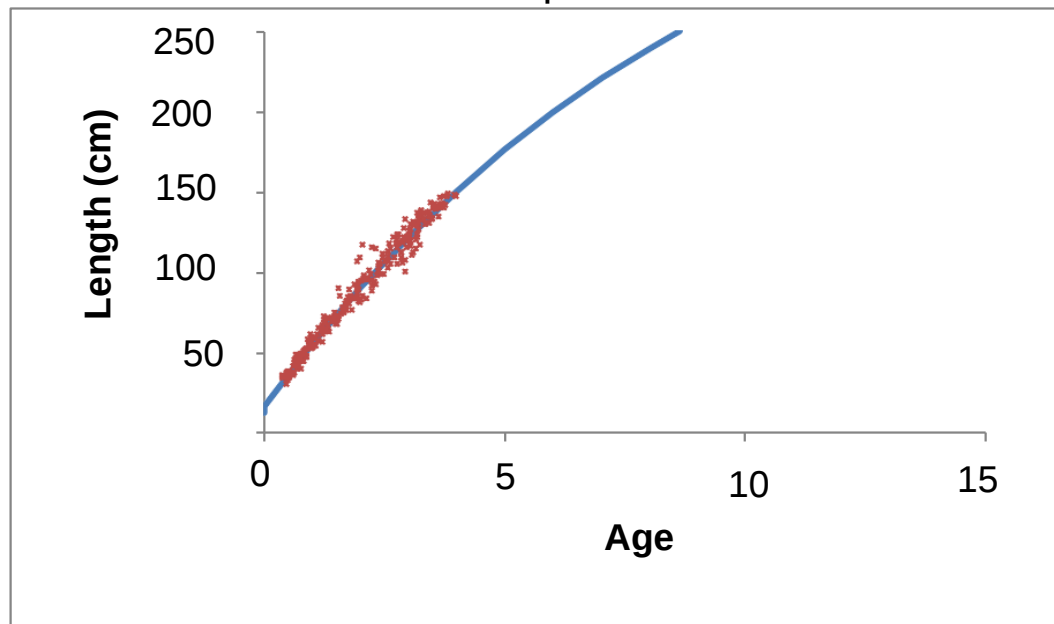
Outline

- Motivation, models and data
- Some current diagnostics tests
- Diagnostics innovation, adoption and diffusion

Motivation, models and data

- We want to “validate” quantities of management interest (e.g. SSB and F) but we can’t to do that.
- Such quantities are not observed, they are essentially an extrapolation of the data used in the model.
- We have to assume the model is correct to extrapolate.

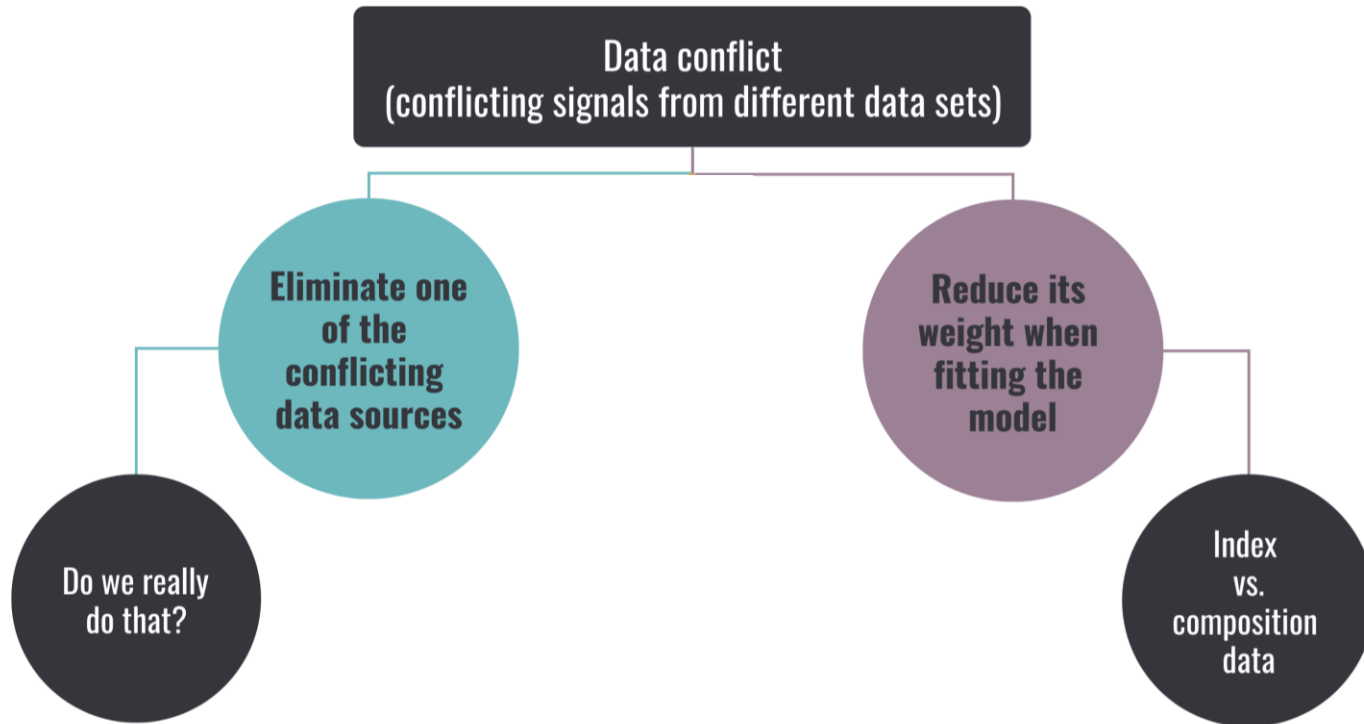
von
Bertalanffy



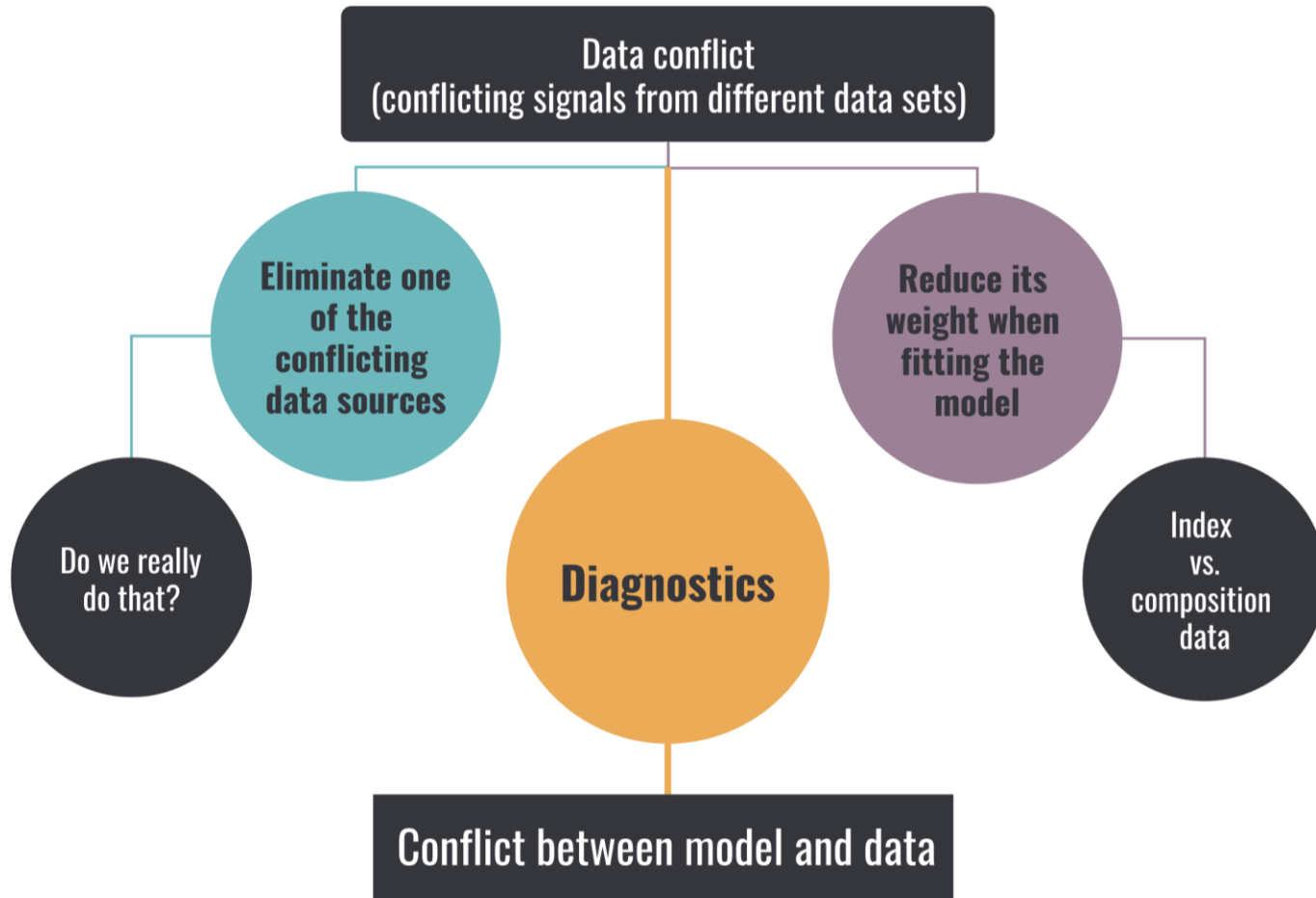
Model misspecification is inevitable

- 1 Incorrect specification of a model parameter
- 2 Using an incorrect model structure
- 3 Incorrect specification of the likelihood functions
- 4 Incorrect specification of the observation model
- 5 Incorrect specification of the system dynamic model
- 6 Ignoring process variability

Motivation, models and data



Motivation, models and data





ELSEVIER

Contents lists available at [ScienceDirect](#)

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres



Full length article

Can diagnostic tests help identify model misspecification in integrated stock assessments?

Felipe Carvalho^{a,b,*}, André E. Punt^c, Yi-Jay Chang^d, Mark N. Maunder^{e,f}, Kevin R. Piner^g

^a University of Hawaii, Joint Institute for Marine and Atmospheric Research, Honolulu, HI 96822, USA

^b NOAA, National Marine Fisheries Service, Pacific Islands Fisheries Science Center, Honolulu, HI 96818, USA

^c School of Aquatic and Fishery Sciences, University of Washington, Seattle, WA 98195, USA

^d Institute of Oceanography, National Taiwan University, Taipei 10617, Taiwan

^e Inter-American Tropical Tuna Commission, La Jolla Shores Drive, La Jolla, CA 92037, USA

^f Center for the Advancement of Population Assessment Methodology, Scripps Institution of Oceanography, La Jolla, USA

^g NOAA Fisheries, Southwest Fisheries Science Center, La Jolla Shores Dr, La Jolla, CA 92037, USA



Motivation, models and data

- Models must be validated if they are to provide credible and robust advice.
- Validation includes transferring confidence to people not directly involved in model construction.
- Model diagnostics play an important role to establish a base model or an ensemble of candidate models.



Shortfin mako shark



Pacific hake

Fisheries Research 240 (2021) 105959

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishes

 **ELSEVIER**





A cookbook for using model diagnostics in integrated stock assessments

Felipe Carvalho^{a,*}, Henning Winker^{b,1}, Dean Courtney^c, Maia Kapur^d, Laurence Kell^e, Massimiliano Cardinale^f, Michael Schirripa^g, Toshihide Kitakado^h, Dawit Yemaneⁱ, Kevin R. Piner^j, Mark N. Maunder^{k,1}, Ian Taylor^m, Chantel R. Wetzel^m, Kathryn Doeringⁿ, Kelli F. Johnsonⁿ, Richard D. Methot^m

^a NOAA Fisheries, Pacific Islands Fisheries Science Center, Honolulu, HI, United States
^b Joint Research Centre (JRC), European Commission, Ispra, Varese, Italy
^c NOAA Fisheries, Southeast Fisheries Science Center, Panama City Laboratory, Panama City, FL, United States
^d School of Aquatic and Fisheries Sciences, University of Washington, Seattle, WA, United States
^e Center for Environmental Policy, Imperial College London, London, United Kingdom
^f Swedish University of Agricultural Sciences, Department of Aquatic Resources, Institute of Marine Research, Turistgatan, Lysekil, Sweden
^g NOAA Fisheries, Southeast Fisheries Science Center, Miami, FL, United States
^h Department of Marine Biosciences, Tokyo University of Marine Science and Technology, Konan, Minato, Tokyo, Japan
ⁱ Department of Environment, Forestry and Fisheries, Forestry Building, Cape Town, South Africa
^j NOAA Fisheries, Southwest Fisheries Science Center, La Jolla, CA, United States
^k Inter-American Tropical Tuna Commission, La Jolla, CA, United States
^l Center for the Advancement of Population Assessment Methodology (CAPAM), United States
^m NOAA Fisheries, Northwest Fisheries Science Center, Seattle, WA, United States
ⁿ Caelum Research Corporation Under Contract to National Marine Fisheries Service - Northwest Fisheries Science Center, Seattle, WA, United States

Outline

- Motivation, models and data
- **Some current diagnostics tests**
- Diagnostics innovation, adoption and diffusion

Some current diagnostic tests

Goodness-of-fit



**Are the residuals
sufficiently
random?**

Some current diagnostic tests

Goodness-of-fit



**Information source
and structure**



**Are the residuals
sufficiently
random?**

**Any sign of data
conflict?**

**Presence of
production
function?**

**Evidence of
retrospective
patterns?**

Some current diagnostic tests

Goodness-of-fit



**Information source
and structure**



Prediction skill



**Are the residuals
sufficiently
random?**

**Any sign of data
conflict?**

**Presence of
production
function?**

**Evidence of
retrospective
patterns?**

**Model shows
predictive skills?**

Some current diagnostic tests

Goodness-of-fit



**Are the residuals
sufficiently
random?**

**Information source
and structure**



**Any sign of data
conflict?**
**Presence of
production
function?**
**Evidence of
retrospective
patterns?**

Prediction skill



**Model shows
predictive skills?**

Convergence



**Model has
converged to a
global solution?**
**Is the Hessian
positive definite?**
**Are there any
highly correlated
parameters or with
excessively high
variance?**

Some current diagnostic tests

Goodness-of-fit 	Information source and structure 	Prediction skill 	Convergence 	Plausibility 
Are the residuals sufficiently random?	Any sign of data conflict? Presence of production function? Evidence of retrospective patterns?	Model shows predictive skills?	Model has converged to a global solution? Is the Hessian positive definite? Are there any highly correlated parameters or with excessively high variance?	Does the model meet basic plausibility? • adequate diagnostics. • no “surprising” outlier dynamics. • meets prior knowledge about the exploitation history?

Goodness-of-fit



Information source
and structure



Prediction skill



Convergence

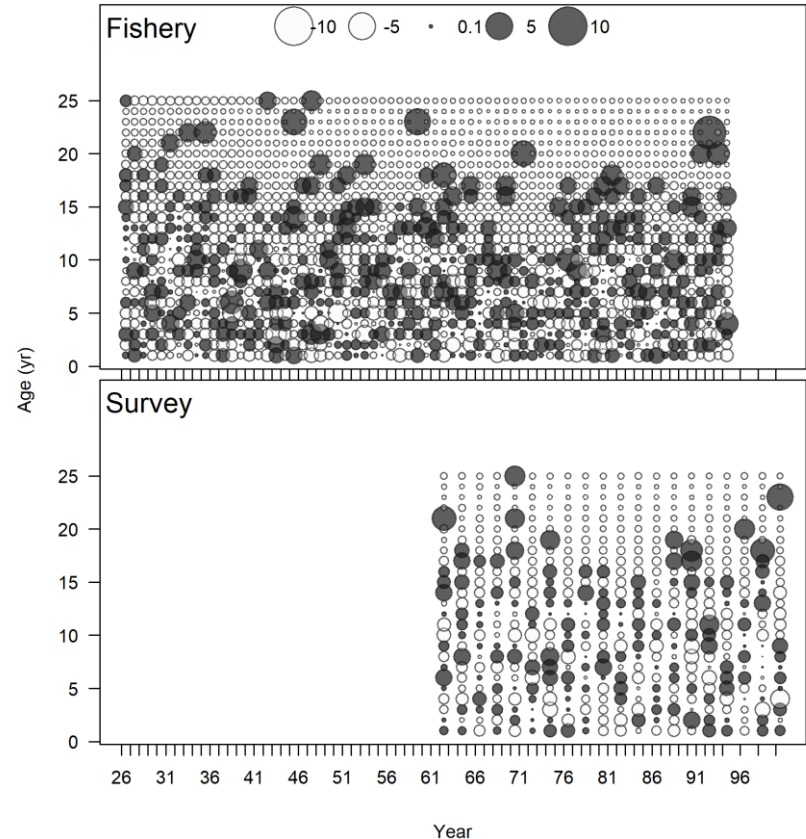


Plausibility



Residual analysis

- Systematic misfit to data should be considered a sign of model misspecification.
- Unacceptable model fits can be detected by either the magnitude or the presence of trends in residuals



Goodness-of-fit



Information source
and structure



Prediction skill



Convergence

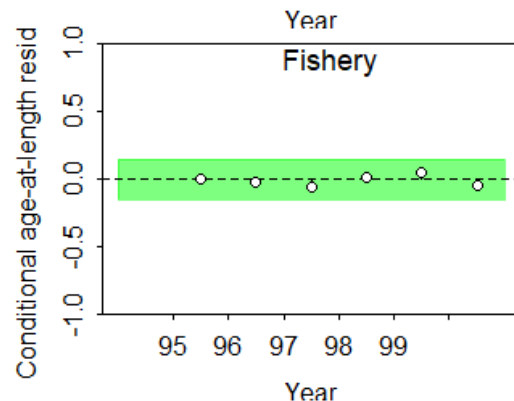
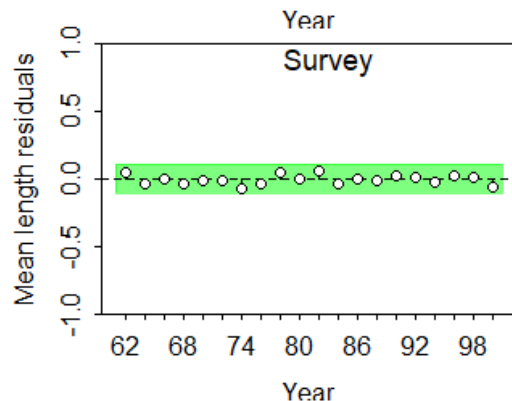
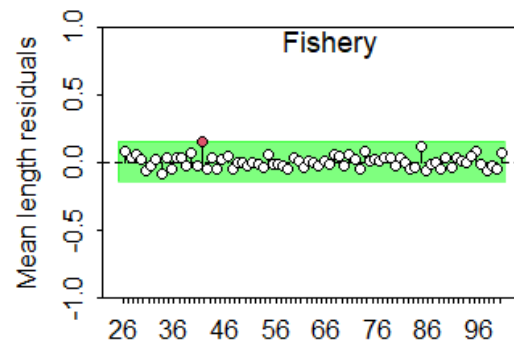
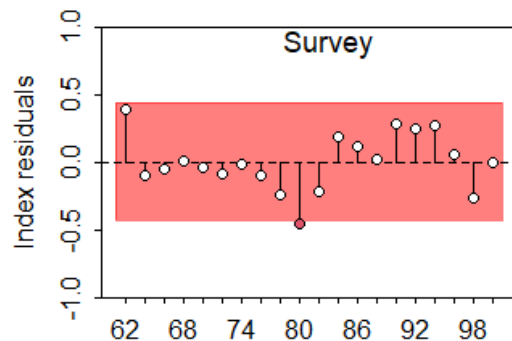


Plausibility



Residual analysis

- Analysts are also interested in whether the probability of being on either side of the median varies with time.
- The runs test is a nonparametric hypothesis test for randomness in a data sequence.



Goodness-of-fit



Information source
and structure



Prediction skill



Convergence

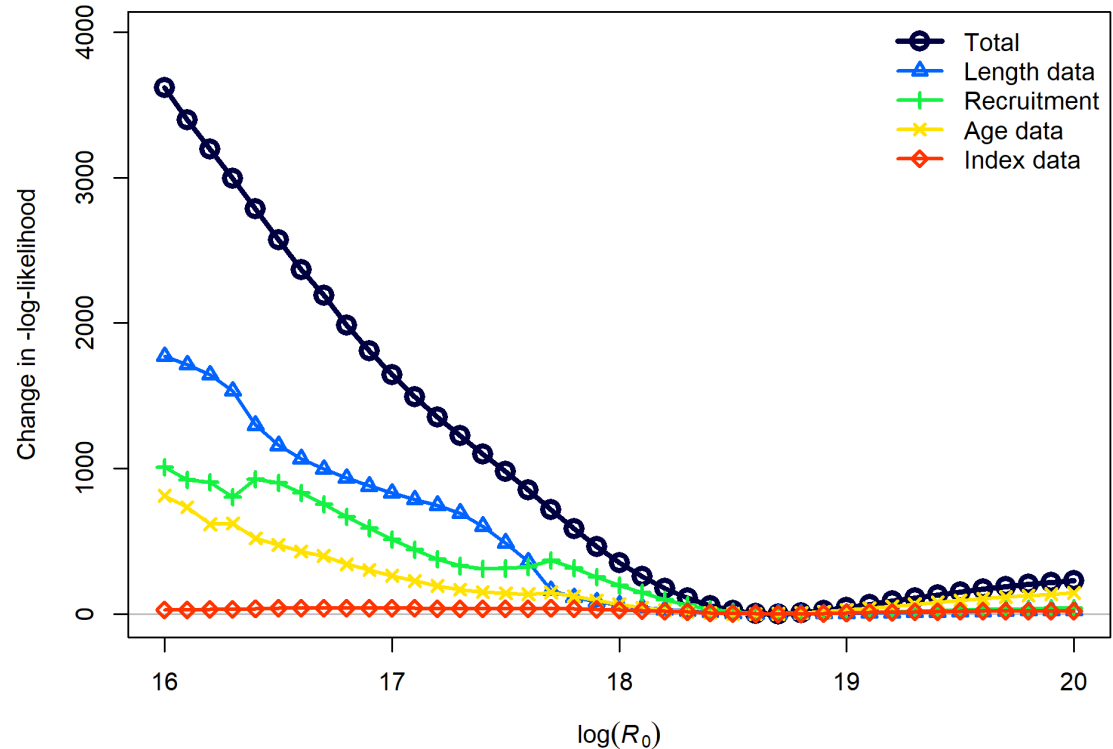


Plausibility



Likelihood profile

- A key model diagnostic developed to identify the influence of information sources on model estimates is the likelihood component profile.



Goodness-of-fit



Information source
and structure



Prediction skill



Convergence



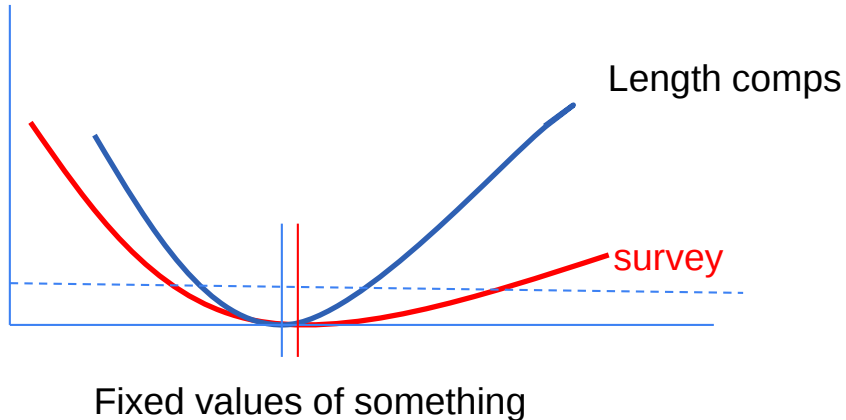
Plausibility



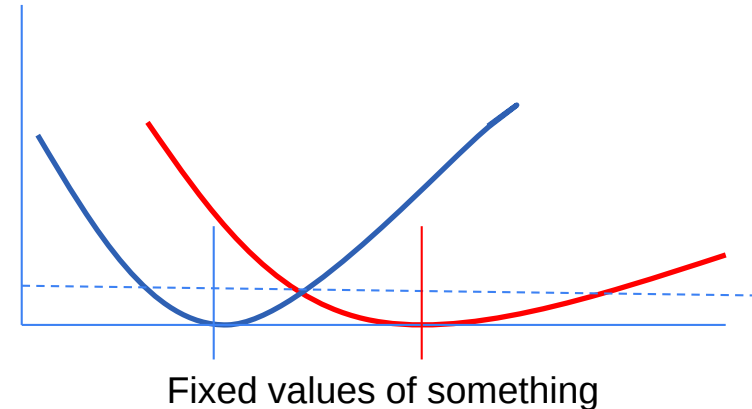
Likelihood profile

- A difference in the location of the minimum negative log-likelihood might suggest either conflict in the data or model misspecification (or both).

No conflict



Most models



Goodness-of-fit



Information source
and structure



Prediction skill



Convergence

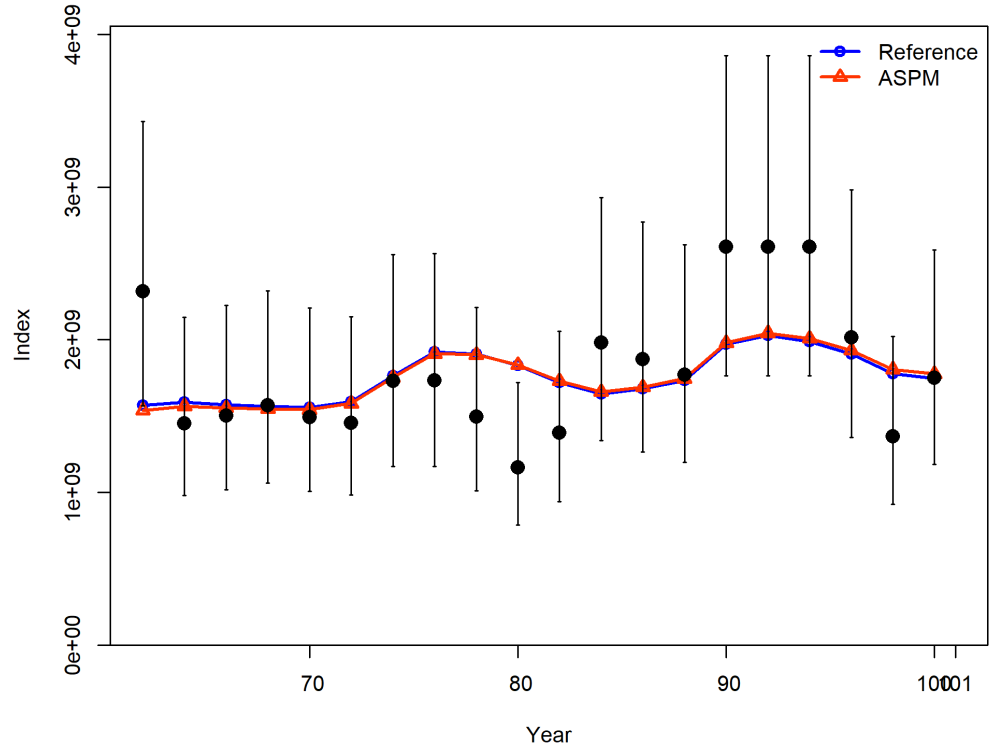


Plausibility



Age-structured production model (ASPM)

- ASPM diagnostic can detect misspecification of key systems-modeled processes that control the shape of the production function.
- This diagnostic evaluates whether the net effect between surplus production and observed catches alone could explain trends in the index of abundance.



Goodness-of-fit



Information source
and structure



Prediction skill



Convergence

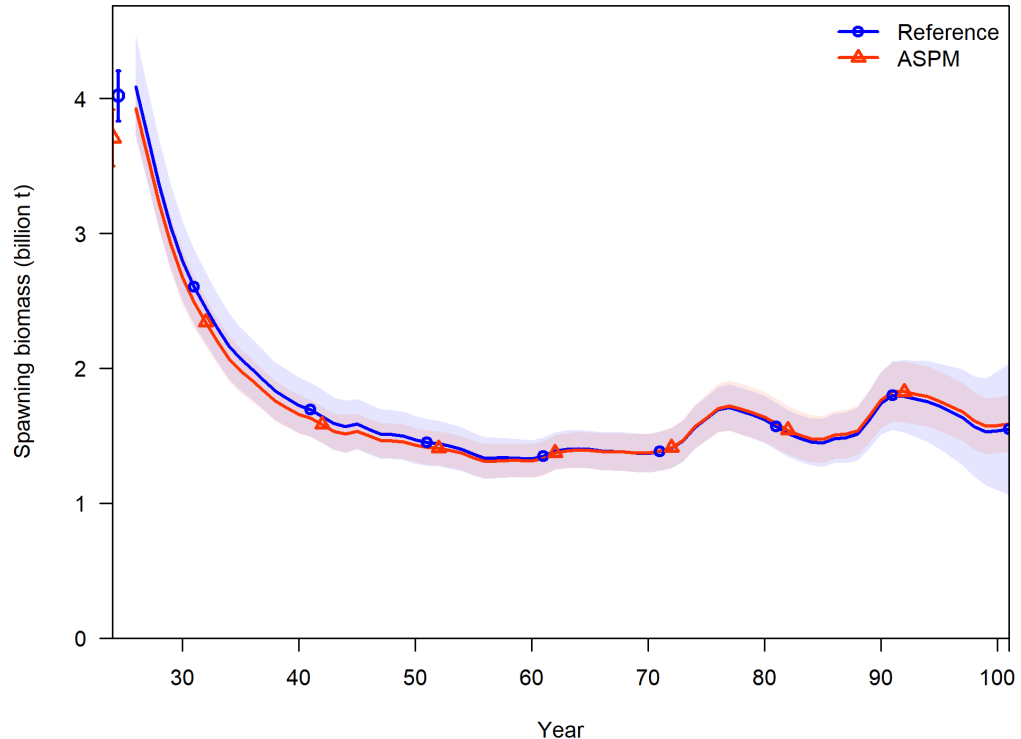


Plausibility



Age-structured production model (ASPM)

- ASPM diagnostic can detect misspecification of key systems-modeled processes that control the shape of the production function.
- This diagnostic evaluates whether the net effect between surplus production and observed catches alone could explain trends in the index of abundance.



Goodness-of-fit



Information source
and structure



Prediction skill



Convergence

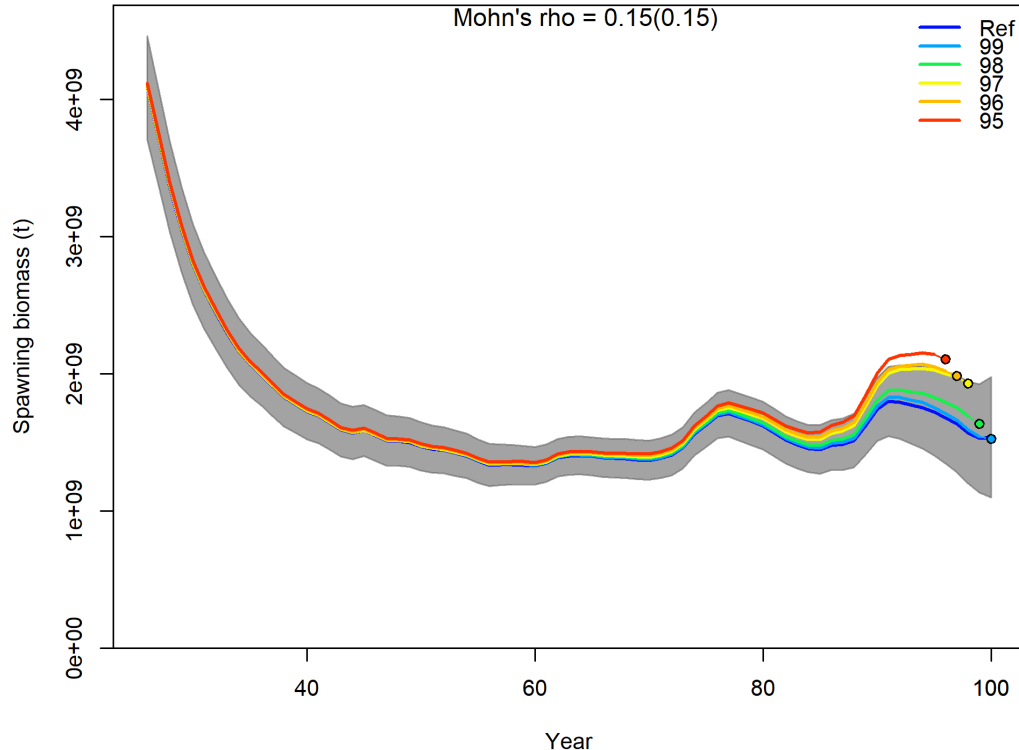


Plausibility



Retrospective analysis

- Commonly used to check the consistency of model estimates as the model is updated with new data in retrospect.
- The most commonly used statistic for retrospective bias is the Mohn's rho.



Goodness-of-fit



Information source
and structure



Prediction skill



Convergence



Plausibility



Hindcasting cross-validation (HCXval)

- Observations are compared to their predicted future values.
- HCXval approach is all about ‘prediction skill’, which is defined as any measure of the accuracy of a forecasted value to the actual observed value.

Fisheries Research 183 (2016) 119–127



ELSEVIER

Contents lists available at [ScienceDirect](#)

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres

Evaluation of the prediction skill of stock assessment using hindcasting

Laurence T. Kell^{a,*}, Ai Kimoto^b, Toshihide Kitakado^c

^a ICCAT Secretariat, C/Corazón de María, 8, 28002 Madrid, Spain

^b Bluefin Tuna Resources Division, National Research Institute of Far Seas Fisheries, Fisheries Research Agency, 5-7-1 Orido, Shimizu, Shizuoka 424-8633, Japan

^c Faculty of Marine Science, Tokyo University of Marine Science and Technology, Department of Marine Biosciences, 5-7, Konan 4, Minato-ku, Tokyo 108-8477, Japan

Goodness-of-fit



Information source
and structure



Prediction skill



Convergence

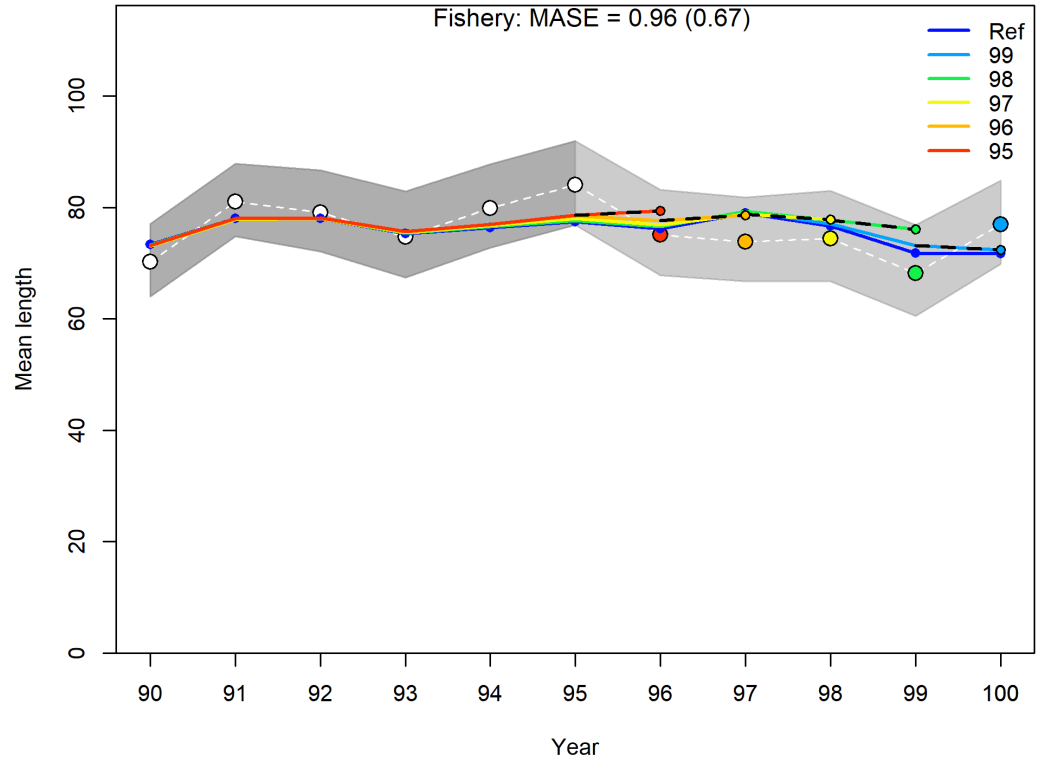


Plausibility



Hindcasting cross-validation (HCXval)

- The HCXval algorithm is similar to that used in the retrospective analysis.
- A robust statistic for evaluating prediction skill is the mean absolute scaled error (MASE).



Goodness-of-fit



Information source
and structure



Prediction skill



Convergence



Plausibility



Hessian and Jitter

- None of the convergence diagnostics alone may be sufficient to demonstrate convergence or the lack of it decisively.
- ✓ Parameters estimated at a bound
- ✓ Small final gradient
- ✓ Hessian is positive definite
- ✓ Correlation matrix
- ✓ Jittering

Goodness-of-fit



Information source
and structure



Prediction skill



Convergence



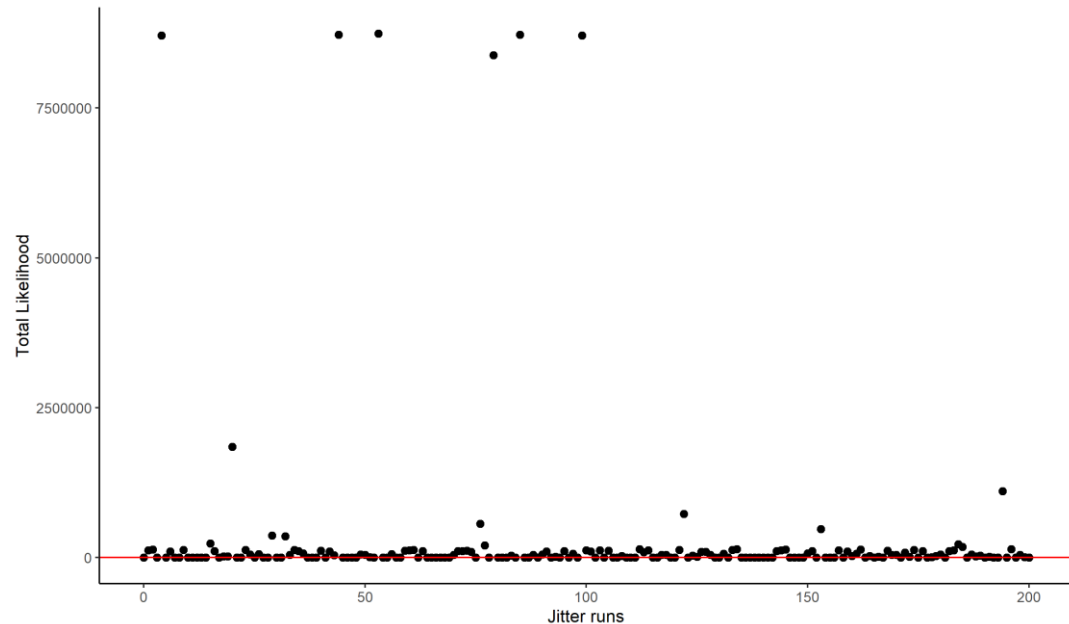
Plausibility



Hessian and Jitter

- None of the convergence diagnostics alone may be sufficient to demonstrate convergence or the lack of it decisively.

- ✓ Parameters estimated at a bound
- ✓ Small final gradient
- ✓ Hessian is positive definite
- ✓ Correlation matrix
- ✓ Jittering



Goodness-of-fit



Information source
and structure



Prediction skill



Convergence



Plausibility



- *Plausibility*: A term often used but seldom defined in stock assessment.



Goodness-of-fit



Information source
and structure



Prediction skill



Convergence



Plausibility



INTER-AMERICAN TROPICAL TUNA COMMISSION

SCIENTIFIC ADVISORY COMMITTEE

11TH MEETING

San Diego, California (USA)

11-15 May 2020¹

DOCUMENT SAC-11 INF-F REV

IMPLEMENTING REFERENCE POINT-BASED FISHERY HARVEST CONTROL RULES
WITHIN A PROBABILISTIC FRAMEWORK THAT CONSIDERS MULTIPLE
HYPOTHESES

Mark N. Maunder, Haikun Xu, Cleridy E. Lennert-Cody, Juan L. Valero, Alexandre Aires-da-Silva,
Carolina Minte-Vera



Goodness-of-fit



Information source
and structure



Prediction skill



Convergence



Plausibility



INTER-AMERICAN TROPICAL TUNA COMMISSION

SCIENTIFIC ADVISORY COMMITTEE

11TH MEETING

San Diego, California (USA)

11-15 May 2020¹

DOCUMENT SAC-11 INF-F REV

IMPLEMENTING REFERENCE POINT-BASED FISHERY HARVEST CONTROL RULES
WITHIN A PROBABILISTIC FRAMEWORK THAT CONSIDERS MULTIPLE
HYPOTHESES

Mark N. Maunder, Haikun Xu, Cleridy E. Lennert-Cody, Juan L. Valero, Alexandre Aires-da-Silva,
Carolina Minte-Vera



- $W(\text{Expert})$: Expert opinion, assigned “a-priori”, without consideration of model fit.
- $W(\text{Convergence})$: Model convergence criteria of the estimation algorithm.
- $W(\text{Fit})$: The fit of the model to the data.
- $W(\text{Plausible parameters})$: The plausibility of the estimates of the parameters representing the hypothesis.
- $W(\text{Plausible results})$: The plausibility of the model results.
- $W(\text{Diagnostics})$: Reliability of the model based on diagnostics.

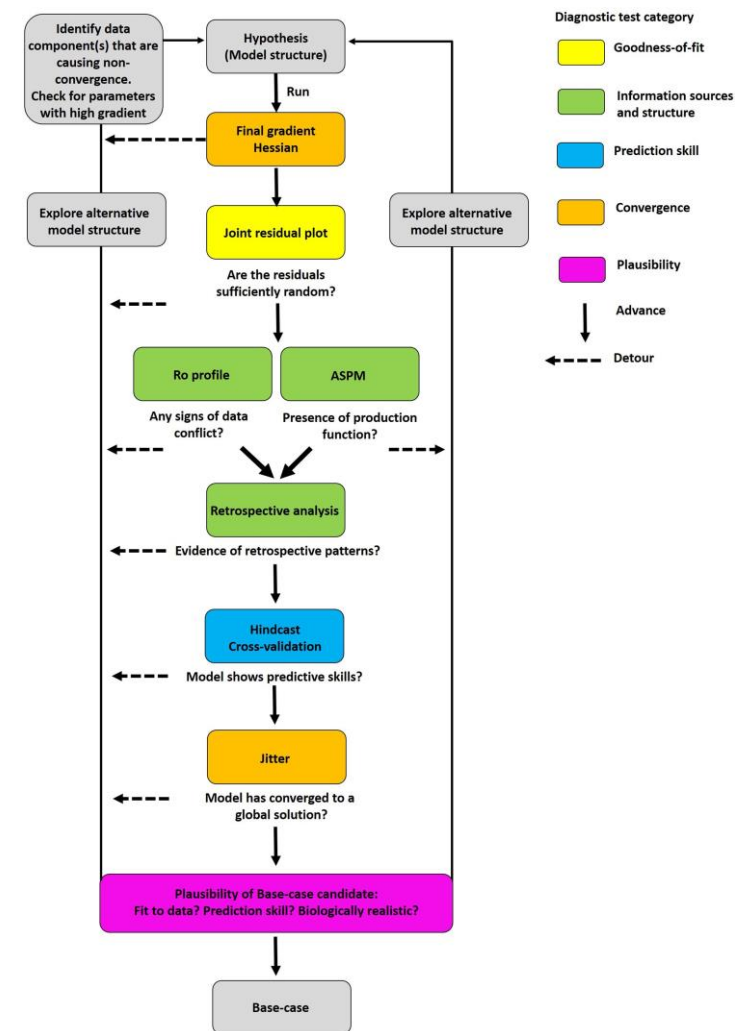
Outline

- Motivation, models and data
- Some current diagnostics tests
- **Diagnostics innovation, adoption and diffusion**

Diagnostics innovation, adoption, and diffusion

✓ Innovation

- Generic process of model development and selection using the presented model diagnostics
- Create the initial model based on the hypothesis and the important data, and decide whether:
 - The optimization was successful
 - The model fits the data;
 - Data has information on trends and scale
 - Estimates are consistent when updated with new data; and
 - The model is able to make future predictions.
- Modify or add additional process as appropriate (diagnostics again and again...)



✓ Adoption



REPORT OF THE 2021 MEETING OF THE WORKING GROUP ON STOCK ASSESSMENT METHODS (WGSAM) *(Online, 5-10 May 2021)*

“The Group recommends the SCRS routinely apply objective criteria for model plausibility for all ICCAT stock assessments that are intended for management (e.g., TAC) advice. These criteria shall be based on best practice in using model diagnostics for evaluating (1) model convergence, (2) fits to the data, (3) model consistency (e.g., retrospective patterns) and (4) prediction skill, as well as biological plausibility criteria. The Group recommends the model diagnostics applied are similar, but not limited to those described in Carvalho *et al.*, (2021).”

Diagnostics innovation, adoption, and diffusion

✓ Adoption



**NOAA
FISHERIES**

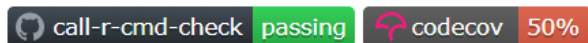
Guidance on model diagnostics for stock assessments in the United States

NMFS Assessment Methods Working Group (AMWG)

- Ongoing effort to describe use of diagnostics in a U.S. regional context, document methods, provide recommendations, and facilitate training.
- Topics include
 - Model convergence
 - Goodness of fit
 - Sensitivity analyses
 - Retrospective analyses
- NOAA Technical Memorandum in prep.

✓ Diffusion

r4ss: R code for Stock Synthesis



Stock Synthesis is a fisheries stock assessment model written in ADMB by Rick Methot. The Stock Synthesis software and many other associated materials are available on the NOAA Virtual Laboratory at <https://vlab.noaa.gov/web/stock-synthesis/home>. The r4ss package is a collection of R functions for interacting with Stock Synthesis. It is based on the original work of Ian Stewart begun around 2005 and released as an open source R package in 2009. The package has a long list of authors and has benefited from a large community of users making suggestions and reporting issues.

Code available on this website comes with no warranty or guarantee of accuracy. It merely represents an ongoing attempt to integrate output plotting, statistics and diagnostics. It is absolutely necessary that prior to use with a new application, the user checks the output manually to verify that there are no plotting or statistical bugs which could incorrectly represent the output files being analyzed.

Diagnostics innovation, adoption, and diffusion

✓ Diffusion

ss3diags **1.0.8.9004**ReferenceChangelog

ss3diags

Build Status

The R package `ss3diags` enables users to apply advanced diagnostics to evaluate a Stock Synthesis model. Diagnostics include residual analyses, hindcasting and cross-validation techniques, and retrospective analyses. Functions also allow users to reproduce the key model diagnostics plots that presented in the paper ‘A Cookbook for Using Model Diagnostics in Integrated Stock Assessments’.

A handbook with detailed [User guidelines for Advanced Model Diagnostics with ss3diags](#) is currently being finalized.

In addition, the ss3diags Github repository provides fully commented step-by-step R recipes on how to:

- [Do log-likelihood profiling for R0](#)
- [Run the ASPM diagnostic](#)
- [Conduct iterative hindcasts for retrospective analysis with forecasts](#)
- [Do Jitter tests](#)

with Stock Synthesis by making use of a comprehensive collection of R functions available in the R package `r4ss`.

Installation

ss3diags is not currently supported on CRAN. You can install the development version of ss3diags from [GitHub](#) with:

```
# install.packages("remotes")
remotes::install_github("PIFSCstockassessments/ss3diags")
```

Applying ss3diags for Model Diagnostics

License

EUPL

Community

[Contributing guide](#)

[Code of conduct](#)

Citation

[Citing ss3diags](#)

Developers

Henning Winker
Author, maintainer

Felipe Carvalho
Author

Massimiliano Cardinale
Author

Laurence Kell
Author

Megumi Oshima
Author

Eric Fletcher
Author

Dev status

 R-CMD-check **passing**

Diagnostics innovation, adoption, and diffusion

✓ Diffusion

r4ss: R code for Stock Synthesis

Call r-cmd-check [passing](#) [codecov](#) [50%](#)

Stock Synthesis is a fisheries stock assessment model written in ADMB by Rick Methot. The Stock Synthesis software and many other associated materials are available on the NOAA Virtual Laboratory at <https://vlab.noaa.gov/web/stock-synthesis/home>. The r4ss package is a collection of R functions for interacting with Stock Synthesis. It is based on the original work of Ian Stewart begun around 2005 and released as an open source R package in 2009. The package has a long list of authors and has benefited from a large community of users making suggestions and reporting issues.

Code available on this website comes with no warranty or guarantee of accuracy. It merely represents an ongoing attempt to integrate output plotting, statistics and diagnostics. It is absolutely necessary that prior to use with a new application, the user checks the output manually to verify that there are no plotting or statistical bugs which could incorrectly represent the output files being analyzed.

ss3diags [10.0.004](#) [Reference](#) [Changelog](#)

ss3diags

Build Status

The R package `ss3diags` enables users to apply advanced diagnostics to evaluate a Stock Synthesis model. Diagnostics include residual analysis, hindcasting and cross-validation techniques, and retrospective analysis. Functions also allow users to reproduce the key model diagnostics plots that presented in the paper *A Cookbook for Using Model Diagnostics in Integrated Stock Assessments*.

A handbook with detailed *User guidelines for Advanced Model Diagnostics with ss3diags* is currently being finalized.

In addition, the `ss3diags` GitHub repository provides fully commented step-by-step R recipes on how to:

- Do top-level model profiling for R0
- Run the ASPM diagnostic
- Conduct relative hindcasts for retrospective analysis with forecasts
- Do other tests

with Stock Synthesis by making use of a comprehensive collection of R functions available in the R package `r4ss`.

Installation

`ss3diags` is not currently supported on CRAN. You can install the development version of `ss3diags` from GitHub with:

```
# Install packages ("remotes")
remotes::install_github("r4ss/stockassessments/ss3diags")
```

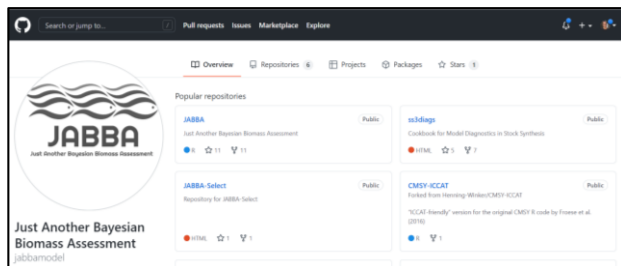
Applying ss3diags for Model Diagnostics

License
EUPL
Community
[Contributing guide](#)
[Code of conduct](#)
Citation
[Citing ss3diags](#)

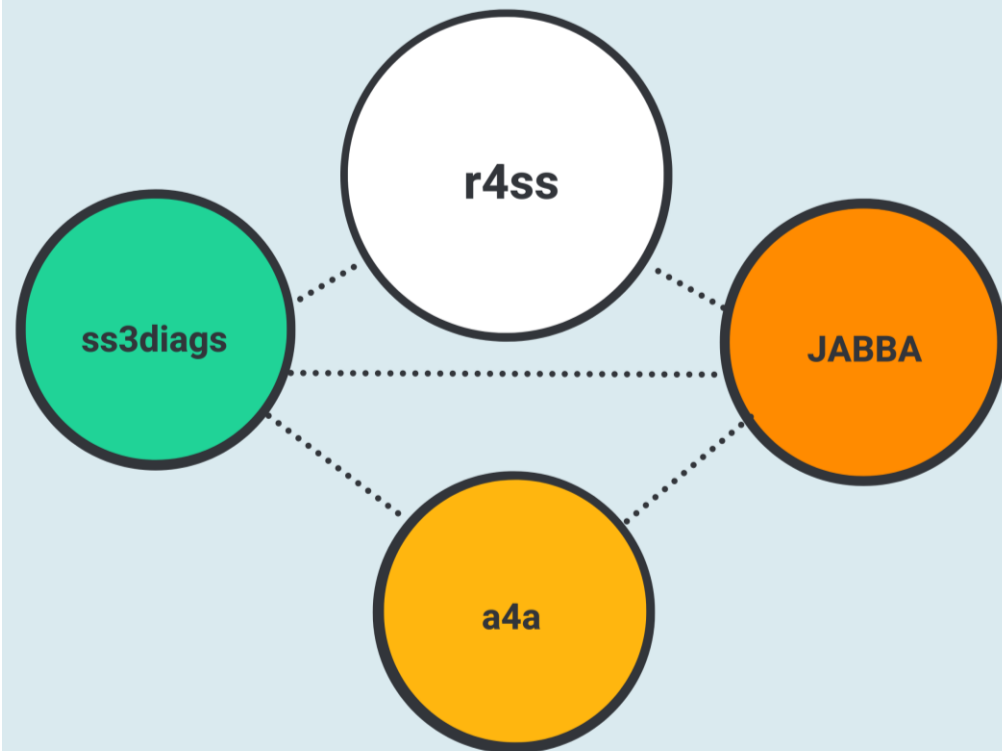
Developers
Hermann Wehrer
Author: maintainer
Felipe Canabarro
Author
Massimiliano Cordone
Author
Laurence Kell
Author
Megumi Oshima
Author
Eric Falcher
Author

Dev status

Call R CMD check [passing](#)



Transferability



Mahalo nui loa!

