Introducing RELAX: An automated pre-processing pipeline for cleaning EEG data - Part 1: Algorithm and application to oscillations

N.W. Bailey [a,b,c,*], M. Biabani [d,1], A.T. Hill [a,e,1], A. Miljevic [a], N.C. Rogasch [d,f,g], B. McQueen [a], O.W. Murphy [a,h], P.B. Fitzgerald [a,b,c]

[a] Central Clinical School Department of Psychiatry, Monash University, Camberwell, Victoria, Australia
[b] School of Medicine and Psychology, The Australian National University, Canberra, ACT, Australia
[c] Monarch Research Institute Monarch Mental Health Group, Sydney, NSW, Australia
[d] The Turner Institute for Brain and Mental Health, School of Psychological Sciences, Monash University, Victoria, Australia
[e] Cognitive Neuroscience Unit, School of Psychology, Deakin University, Melbourne, Victoria, Australia
[f] Discipline of Psychiatry, Adelaide Medical School, University of Adelaide, Adelaide, SA, Australia
[g] Hopwood Centre for Neurobiology, Lifelong Health Theme, South Australian Health and Medical Research Institute (SAHMRI), Adelaide, SA, Australia
[h] Bionics Institute, East Melbourne, VIC 3002, Australia

HIGHLIGHTS

- RELAX is a fully automated EEG cleaning pipeline, freely available on GitHub, and easy to use through a graphical user interface.
- RELAX provided amongst the best performance out of a range of common EEG pre-processing pipelines at cleaning all artifact types.
- RELAX provided high values for the variance explained by the difference between outcomes in common experimental manipulations.

ARTICLE INFO

ABSTRACT

Objective: Electroencephalographic (EEG) data are often contaminated with non-neural artifacts which can confound experimental results. Current artifact cleaning approaches often require costly manual input. Our aim was to provide a fully automated EEG cleaning pipeline that addresses all artifact types and improves measurement of EEG outcomes

Methods: We developed RELAX (the Reduction of Electroencephalographic Artifacts). RELAX cleans continuous data using Multi-channel Wiener filtering [MWF] and/or wavelet enhanced independent component analysis [wICA] applied to artifacts identified by ICLabel [wICA_ICLabel]. Several versions of RELAX were compared using three datasets (N = 213, 60 and 23 respectively) against six commonly used pipelines across a range of artifact cleaning metrics, including measures of remaining blink and muscle activity, and the variance explained by experimental manipulations after cleaning.

Results: RELAX with MWF and wICA_ICLabel showed amongst the best performance at cleaning blink and muscle artifacts while preserving neural signal. RELAX with wICA_ICLabel only may perform better at differentiating alpha oscillations between working memory conditions.

Conclusions: RELAX provides automated, objective and high-performing EEG cleaning, is easy to use, and freely available on GitHub.

Significance: We recommend RELAX for data cleaning across EEG studies to reduce artifact confounds, improve outcome measurement and improve inter-study consistency.

© 2023 International Federation of Clinical Neurophysiology. Published by Elsevier B.V. All rights reserved.

* Corresponding author.
  E-mail address: neil.bailey.research@gmail.com (N.W. Bailey).
[1] These authors contributed equally to the manuscript.

The MATLAB code can be downloaded from: https://github.com/NeilwBailey/RELAX/releases. Instructions for installing and using the pipeline can be found in the GitHub wiki: https://github.com/NeilwBailey/RELAX/wiki. A condition of use of the pipeline is that the version of the pipeline used is referred to as RELAX_[pipeline], for example "RELAX_MWF_wICA" or "RELAX_wICA_ICLabel", that the current paper be cited, as well as the dependencies used, and that authors report the RELAX settings that were planned a priori, as well as the settings that were implemented (if these differ). These dependencies are likely to include: EEGLAB (Delorme and Makeig, 2004), fieldtrip (Oostenveld et al., 2011), the MWF toolbox (Somers et al., 2019), fastICA (Hyvarinen, 1999), wICA (Castellanos and Makarov, 2006), ICLabel (Pion-Tonachini et al., 2019), and PREP (Bigdely-Shamlo et al., 2015).

The application of RELAX to studies of event-related potentials (ERPs) is reported in our companion paper (Bailey et al., 2022).

1. Introduction

Electroencephalography (EEG) allows investigators to non-invasively measure voltage fluctuations generated by the brain. As such, EEG is useful for uncovering functional relationships between neural activity and cognition, as well as for assessing how neural activity differs between healthy and clinical populations. However, the voltage fluctuations detected by EEG electrodes are produced not only by neural activity, but also by non-brain related "artifacts". These artifacts can be biological in origin, for example eye blinks, movements, and muscle activity. Biological artifacts generate electrical potentials that show stereotypical characteristics, making them reasonably easy to identify (Fitzgibbon et al., 2016, Kleifges et al., 2017, Muthukumaraswamy, 2013). Non-biological artifacts are also often present. For example, voltage drift, which can be caused by changes to scalp electrical impedances due to participant sweat or electrical interference; 50 Hz or 60 Hz line noise generated by the alternating current of electrical grids; and channel noise due to disrupted electrical connection between electrodes and the scalp (Pion-Tonachini et al., 2019). Because each EEG electrode records a mixture of neural activity and artifacts with an unknown 'ground-truth', it is difficult to precisely disentangle the neural activity from artifacts when analysing EEG signals (Muthukumaraswamy, 2013). To address this issue, a multitude of methods for pre-processing EEG data have been developed (see Islam et al. (2016) and Ranjan et al. (2021) for reviews, and Barban et al. (2021) and Robbins et al. (2020) for comparisons across a range of cleaning methods). The fundamental aim of artifact cleaning is to reduce the influence of artifacts on the EEG data while leaving signals from neural activity unaltered (Muthukumaraswamy, 2013).

Although existing EEG pre-processing methods seem adequate, there are still several outstanding issues that may adversely affect the results of EEG research. These limitations include:

1) The lack of a single gold-standard method or even best single approach (Barban et al., 2021). A large part of the reason for the lack of a gold-standard method is that, unfortunately, the 'ground-truth' of EEG recordings (real brain signal without any artifacts) is typically unknown (Muthukumaraswamy, 2013). As such, while measures can be used to infer the performance of a cleaning pipeline, researchers cannot *prove* a pipeline leads to perfect signal

extraction. As a result, certainty regarding *the* optimal pipeline is elusive. Simulated data that have artifact patterns does provide the ability for researchers to know the ground-truth signal (the simulated neural activity without artifacts) (Urigüen and Garcia-Zapirain, 2015). Unfortunately, it is likely that simulated data does not match real data in a number of different aspects, including synchronisation of activity to stimuli, synchronisation between electrodes (Urigüen and Garcia-Zapirain, 2015), or synchronisation between blink or muscle artifacts and the neural activity that generates these artifacts. Additionally, researchers have noted that typically, realistic artifacts in simulated data are added from an artifact template obtained by extracting real artifacts from recorded data using an existing artifact cleaning method (Barban et al., 2021, Urigüen and Garcia-Zapirain, 2015). The same artifact cleaning method is then included in the testing of how effectively artifacts are cleaned from the simulated data, creating a potential bias towards that cleaning method (Barban et al., 2021, Urigüen and Garcia-Zapirain, 2015). As such, while the simulated data does provide a ground truth, successful cleaning of this simulated data does not guarantee effective cleaning of actual data, and real data provides the most important final test of any cleaning method (Urigüen and Garcia-Zapirain, 2015).

Additionally, newly-developed pipelines are also typically compared against only a small number of existing methods, which further limits the field's ability to determine an optimal EEG pre-processing pipeline. Similarly, few studies have examined a large range of data cleaning quality metrics simultaneously (Robbins et al., 2020). Demonstrating the superior cleaning of a single artifact type (for example muscle activity) does not necessarily mean the effect of other artifacts (for example blinks) are completely mitigated, so evaluating cleaning efficacy for all artifact types simultaneously is important. Additionally, very few EEG cleaning studies have included an assessment of the most practically important metric - whether cleaning increases the signal-to-noise ratio so that experiments are more likely to detect real effects of interest (Clayson et al., 2021a). While these reasons for the lack of a gold-standard EEG cleaning pipeline are reasonable, the absence of consistent cleaning methods across studies limits our ability to compare findings across different studies, as cleaning methods can vary considerably from study to study. This inconsistency has practical consequences, with research demonstrating variability in experimental results when different cleaning pipelines have been applied even to the same dataset (Robbins et al., 2020).

2) The majority of currently available EEG pre-processing pipelines are only partially automated, leaving results vulnerable to potential bias or experimenter error/subjectivity, as well as being time intensive (and often boring!). The lack of full automation also means that significant technical expertise is required to pre-process the EEG data, increasing the barrier for entry to EEG research and decreasing the potential for student projects to provide meaningful contributions.

3) While fully automated pre-processing methods are available (for example, The Harvard Automated Processing Pipeline for Electroencephalography [HAPPE]; Gabard-Durnam et al. (2018)), in our experience, these methods often do not match our expert judgements, either over or under-rejecting data, leading to reductions in signal (an issue that is demonstrated in the results of the current study).

4) While current EEG pre-processing methods seem to be effective for most data, we have found that some data are not effectively cleaned by existing approaches. In fact, the inspi-

ration for the current work was our observation that using a traditional independent component analysis (ICA) subtraction approach, some data had blink artifacts remaining after cleaning, while other data were over-cleaned, leaving an inverted blink signal in the EEG trace. This effect has also been demonstrated in previous research (Dimigen, 2020).

5) Within cleaning pipelines, there are often many parameters available for adjustment and no clear indication of which parameter settings are optimal. This can lead to confusion and wasted exploratory time for researchers. It also poses the issue that multiple analyses could be undertaken to obtain the desired result. If a well-intentioned researcher assumes an effect is likely, it might seem sensible to apply multiple EEG cleaning approaches to ensure null results were not produced by inferior data cleaning, and to only report the expected positive results. Researchers may do this without realising their results might simply be dependent on the cleaning approach used, reducing the replicability of their results and potential producing false positive results.

To address these issues, we have developed a new EEG pre-processing pipeline by combining and adapting pre-existing approaches and iteratively testing all cleaning parameters in an attempt to find the optimal artifact reduction approach which concurrently preserves neural signal. We have called this pipeline RELAX (short for "Reduction of Electroencephalographic Artifacts"). The pipeline removes extreme outlying electrodes and periods using objective methods derived from previous research. These methods were also subjected to extensive informal testing to maximise outcome metrics, and to ensure a match with our expert judgement. The key artifact reduction components of RELAX are Multi-channel Wiener Filters (MWF) (Borowicz, 2018, Somers et al., 2018) and/or wavelet enhanced ICA (wICA) (Castellanos and Makarov, 2006) applied to artifacts detected by the machine learning algorithm 'ICLabel' (Pion-Tonachini et al., 2019). These two approaches are implemented to reduce blinks, muscle activity, horizontal eye movement, voltage drift, and atypical artifacts as effectively as possible while preserving neural signal. The pipeline is implemented in EEGLAB (Delorme and Makeig, 2004), and is fully automated, while also being modular so that the pipeline can be easily adapted for optimal cleaning for a user's intended application (the pipeline is freely available on GitHub at https://github.com/NeilwBailey/RELAX/releases). In the following sections we present the results of the formal comparisons across multiple versions of the RELAX pipeline and six commonly used alternative pipelines, focused on data that is typically examined using the power of neural oscillations as an outcome measure. We tested each pipeline in its ability to clean multiple artifact types, as well as its ability to provide data that maximised the detection of experimental effects of interest in a large heterogenous EEG dataset. We also undertook tests of cleaning efficacy in two independent datasets with different recording parameters to ensure generalisability. Additional comparisons between different versions of the RELAX pipeline (with specific parameters varied) are provided in the Supplementary Materials (sections 5–7). Additionally, our companion article explores the application of RELAX to data analysed using event-related potentials (ERPs) as an outcome measure (Bailey et al., 2022).

2. Methods

2.1. RELAX pipeline

The RELAX pipeline is implemented on raw continuous multi-channel EEG data in EEGLAB format, and uses EEGLAB (Delorme

and Makeig, 2004) and fieldtrip (Oostenveld et al., 2011) functions, implemented within MATLAB. The pipeline outputs cleaned continuous data, with extreme outlying periods and channels removed, referenced to the robust average reference (Bigdely-Shamlo et al., 2015) (note that other re-referencing options can be implemented on this cleaned data). Both MWF and wICA cleaning methods are used in the RELAX pipeline (with the ability to use only one of these methods available for selection in the graphical user interface, and recommendations for specific use cases provided in the discussion). In the following, we provide a narrative description of the RELAX pipeline, with a fully detailed explanation in the Supplementary Materials (section 2, pages 5–10).

MWF has recently been proposed as an effective method for cleaning EEG data (Borowicz, 2018, Somers et al., 2018). The MWF toolbox uses a template of time windows of the continuous EEG data that are identified as showing artifacts to determine artifact patterns in the data that should be reduced, and periods identified as clean data to determine the patterns which should be preserved (Somers et al., 2018). These templates of clean data periods and artifact periods are then transformed into covariance matrices (Borowicz, 2018, Somers et al., 2018). Using the template of clean and artifact periods, a low-rank approximation of the artifact is constructed based on the generalised eigenvalue decomposition (Somers et al., 2018). This approximation contains both spatial and spectral information (Somers et al., 2018). The multi-channel artifact signal estimate is produced by obtaining a solution to the equation for the mean squared error between the artifact and the clean data (which are both included in the equation), with the optimal solution reflecting the minimum mean squared error (Somers et al., 2018). The calculation of this minimum mean squared error provides the artifact signal estimate value (Somers et al., 2018). This artifact signal estimate produces a spatial filter, which is extended into the temporal dimension by applying the spatial filter to each channel with a 'delay period' – a positive and negative time lag from each timepoint (typically 5–10 samples in each direction), which turns the MWF into a finite impulse response filter as well as a spatial filter, which cleans the data based on both spatial and temporal information (Somers et al., 2018). For a more detailed explanation, see Somers et al. (2018). Because the MWF captures both spatial and temporal information in its estimate of the artifact signal, the approach is highly effective at removing artifacts even if they are non-stationary (unlike ICA which only models stationary information) (Borowicz, 2018, Somers et al., 2018). Previously, MWF cleaning has been implemented within the MWF toolbox by manual identification of artifacts (Somers et al., 2019). This is time consuming and vulnerable to inconsistency, mistakes, and subjective decisions, so we have automated the template construction within the RELAX pipeline.

Although MWF cleaning has the aforementioned advantages, it requires a template of both artifact and clean data periods. As such, it cannot address artifacts that are consistent through an entire EEG recording (for example, persistent muscle activity in temporal electrodes or line noise). To resolve this limitation, we tested whether combining wICA and MWF together would address the limitations of each technique when applied separately. wICA has been suggested to address some of the limitations of a traditional ICA subtraction artifact correction method (Castellanos and Makarov, 2006, Inuso et al., 2007). Traditional ICA subtraction methods iteratively models signals of putative source components underlying EEG data to derive source signals that are maximally independent (Delorme and Makeig, 2004). Artifactual components are (mostly) independent of neural activity, so by using ICA, researchers can subtract components identified as artifacts, leaving behind only components that reflect neural activity (Delorme et al., 2007). Scalp level EEG data is then reconstructed, resulting in a

signal reflective of neural activity with artifact influences minimised (Delorme et al., 2007). However, ICA decomposition is adversely affected by large amplitude artifacts, the presence of which leads to components with a mixture of neural activity and artifact (Akhtar et al., 2012, Anders et al., 2020). Additionally, ICA is never a perfect model, so even without large amplitude artifacts ICA components typically contain a mix of neural activity and artifact (Castellanos and Makarov, 2006, Inuso et al., 2007). Due to this mixing, subtraction of putative artifact components also removes neural activity, while keeping putative artifact components fails to clean the artifact (Akhtar et al., 2012). To remedy these limitations, wICA uses wavelet thresholding of ICA components to detect the oscillatory frequencies with the largest influence on the time course of an ICA component (Castellanos and Makarov, 2006). wICA then subtracts wavelet coefficients above a certain threshold (assumed to reflect artifacts) from the component, leaving behind a component with only the smaller amplitude (presumably neural) activity (Castellanos and Makarov, 2006). As with the ICA subtraction method, this artifact reduced component data is then reconstructed into scalp level EEG data. Previous research has commonly applied wICA to all components (Castellanos and Makarov, 2006). However, our experience indicated that applying wICA to all components reduced considerable amounts of the neural signal as well as the artifacts (which can be seen in our results). To prevent this over-cleaning, RELAX applies wICA cleaning only to components that are identified as artifacts by the machine learning algorithm ICLabel (Pion-Tonachini et al., 2019).

Now that we have introduced the techniques underpinning RELAX, we provide below an abbreviated description of the steps undertaken by the RELAX pipeline, with the full details provided in the Supplementary Materials (section 2, pages 5–10). RELAX implements the following:

1) removes extreme outlying electrodes and time periods using methods adapted from previous research which exactly matched our expert judgement (after extensive informal testing). These methods were based on the median and Median Absolute Deviation (MAD) where possible, as this approach produces thresholds that are robust against the effect of more extreme outliers (Alday and van Paridon, 2021);
2) RELAX then performs three separate sequential MWF cleaning of: i) muscle activity, ii) blinks, and iii) horizontal eye movements and drift;
3) Finally, RELAX re-references data to the robust common average (Bigdely-Shamlo et al., 2015), and reduces remaining artifacts using wICA, applied only to independent components identified as artifacts by the machine learning algorithm, ICLabel, which has been trained on large datasets to match expert identification of artifact components (Pion-Tonachini et al., 2019). An overview of the specific steps included in our recommended version of the RELAX pipeline is provided in Fig. 1, and an abbreviated explanation of the full methods are provided in the following section.

2.1.1. Filtering and rejection of extreme outliers

Firstly, a fourth order acausal Butterworth bandpass filter was applied from 1 to 80 Hz and a second order acausal Butterworth notch filter was applied from 47 to 53 Hz (this can be easily adapted for 60 Hz line noise). Note that ICA performs best after data have been filtered with 1 Hz high-pass filters (Winkler et al., 2015), our experience indicates MWF also performs best after 1 Hz high-pass filtering, and filtering out < 1 Hz data typically does not affect oscillation measures of interest. However, high-pass filter settings for ERPs likely need to be < 0.3 Hz, with the exact optimal high-pass filter setting still under debate (Maess et al., 2016,

Rousselet, 2012, Tanner et al., 2016). We address the issue of using RELAX with high-pass filtering appropriate for ERPs in our companion article (Bailey et al., 2022). After filtering, a multi-step process was used to identify and remove bad electrodes. Firstly, the "findNoisyChannels" function from the 'EEG Preprocessing Pipeline' (PREP) was applied to provide an initial rejection of bad electrodes (Bigdely-Shamlo et al., 2015). Then, although RELAX applied MWF and wICA cleaning to continuous data, data were first segmented to identify artifact periods (which were then marked in the continuous data). To implement this, data were split into 1 second epochs with a 0.5 second overlap. Within these epochs, RELAX identified: 1) extreme outlying amplitudes unlikely to reflect data periods with retrievable brain activity, which were identified using MAD from the median, providing robust estimates even with extreme outliers (Alday and van Paridon, 2021); 2) extreme drift; 3) extreme kurtosis; 4) extremely improbable voltage distributions; and finally 5) log-power log-frequency slopes indicative of muscle activity (Fitzgibbon et al., 2016) were used to mark electrodes for rejection if > 5 % of an electrode's data showed contamination with muscle activity. A limit of removing 20 % of electrodes was imposed (a setting which can be adjusted by the user), and if > 20 % of electrodes were marked for rejection, electrodes were ranked in order of the total number of epochs showing extreme artifacts and only the worst 20 % were rejected. Following the electrode rejection step, the same extreme outlier detection approaches were used to mark extreme periods in the EEG data for exclusion. These periods were marked with "not a number" (NaN) within the MWF template. The MWF toolbox ignores these periods when constructing cleaning templates, so that the MWF cleaning focuses on artifacts that are likely to contain retrievable neural activity, rather than periods that may only reflect extreme artifacts without the potential to retrieve underlying neural activity (Somers et al., 2019). These extreme outlier periods were also rejected from the data completely prior to the ICA decomposition for the same reason. Both electrode rejection and extreme outlying period detections were transferred from the epoched data back to the continuous data, which was cleaned in the following steps.

2.1.2. Multi-channel Wiener filtering

Following the extreme outlier rejection steps, three sequential MWF cleaning steps were implemented to address (in chronological order): 1) muscle activity, then 2) blink activity, then 3) both horizontal eye movement and drift activity together. Note that for some applications, skipping straight to wICA might be preferred, which is easily implemented in RELAX (these applications are explored in the discussion). Firstly, to create a template for the detection of muscle activity for MWF cleaning, the data were separated into 1 s epochs with a 500 ms overlap. For each channel, epochs affected by muscle activity were detected using a log-power log-frequency slope threshold of > -0.59, a threshold chosen because very little data within EEG recordings taken from people who have had their muscles pharmacologically paralysed show slopes above this threshold (Fitzgibbon et al., 2016). A more detailed description of the empirical basis for this threshold (and for alternative more liberal or stringent thresholds) is provided in the Supplementary Materials (page 10). A full EEG data length 1D template of muscle artifact periods (marked as 1) and clean periods (marked as 0) was constructed, and the MWF cleaning was applied using this template.

Secondly, to detect blinks, a copy of the data were bandpass filtered using a fourth order Butterworth filter from 1-25 Hz (which focused on this data on blink relevant frequencies). We then averaged pre-specified blink affected electrodes ('FP1'; 'FPZ'; 'FP2'; 'AF3'; 'AF4'; 'F3'; 'F1'; 'FZ'; 'F2'; 'F4' – this list can be easily adapted to match a researcher's EEG files – we recommend only including electrodes that contain large and obvious blink artifacts, ideally
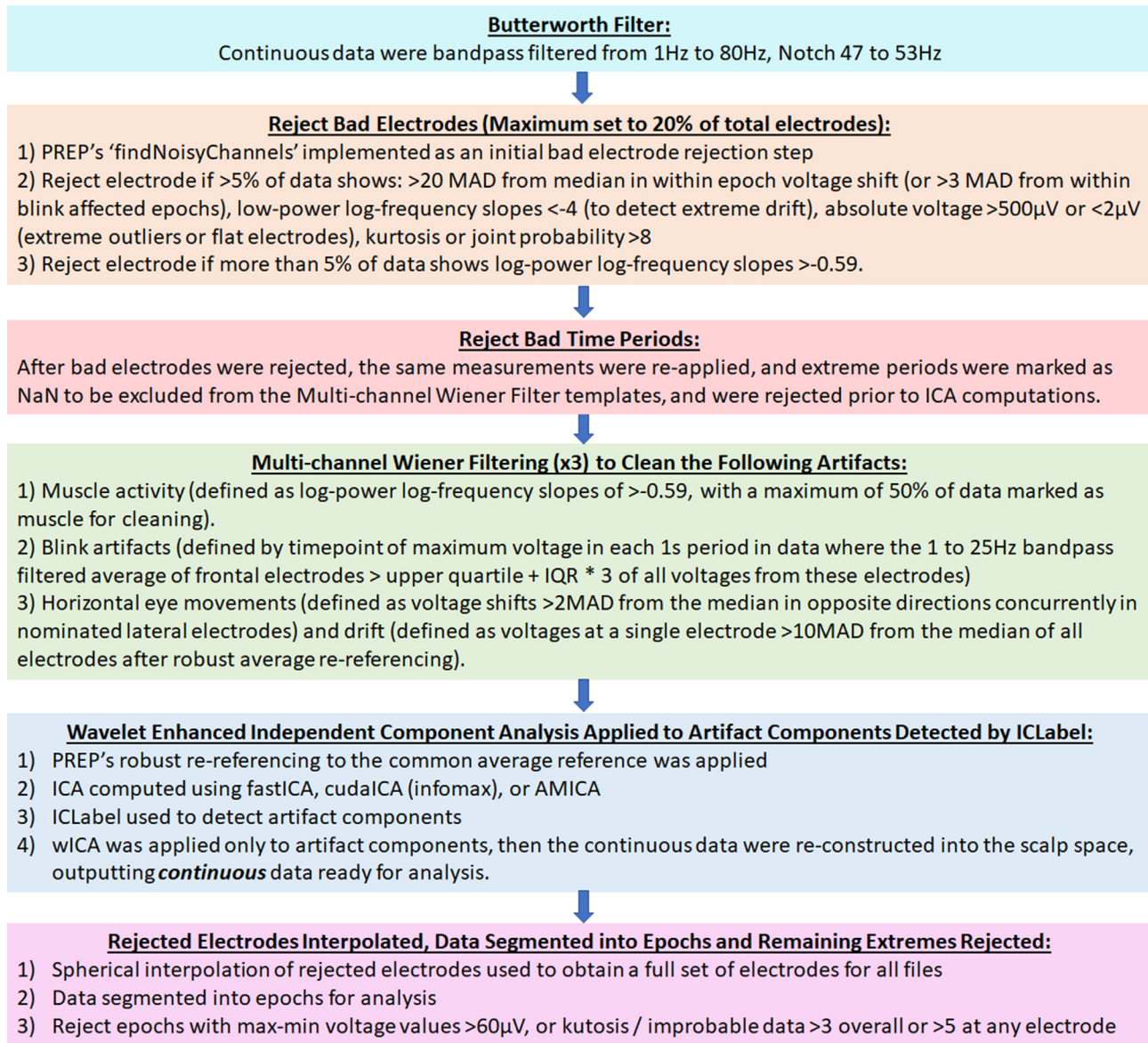
N.W. Bailey, M. Biabani, A.T. Hill et al.

Butterworth Filter:
Continuous data were bandpass filtered from 1Hz to 80Hz, Notch 47 to 53Hz

Reject Bad Electrodes (Maximum set to 20% of total electrodes):
1) PREP's 'findNoisyChannels' implemented as an initial bad electrode rejection step
2) Reject electrode if >5% of data shows: >20 MAD from median in within epoch voltage shift (or >3 MAD from within blink affected epochs), low-power log-frequency slopes <-4 (to detect extreme drift), absolute voltage >500μV or <2μV (extreme outliers or flat electrodes), kurtosis or joint probability >8
3) Reject electrode if more than 5% of data shows log-power log-frequency slopes >-0.59.

Reject Bad Time Periods:
After bad electrodes were rejected, the same measurements were re-applied, and extreme periods were marked as NaN to be excluded from the Multi-channel Wiener Filter templates, and were rejected prior to ICA computations.

Multi-channel Wiener Filtering (x3) to Clean the Following Artifacts:
1) Muscle activity (defined as log-power log-frequency slopes of >-0.59, with a maximum of 50% of data marked as muscle for cleaning).
2) Blink artifacts (defined by timepoint of maximum voltage in each 1s period in data where the 1 to 25Hz bandpass filtered average of frontal electrodes > upper quartile + IQR * 3 of all voltages from these electrodes)
3) Horizontal eye movements (defined as voltage shifts >2MAD from the median in opposite directions concurrently in nominated lateral electrodes) and drift (defined as voltages at a single electrode >10MAD from the median of all electrodes after robust average re-referencing).

Wavelet Enhanced Independent Component Analysis Applied to Artifact Components Detected by ICLabel:
1) PREP's robust re-referencing to the common average reference was applied
2) ICA computed using fastICA, cudaICA (infomax), or AMICA
3) ICLabel used to detect artifact components
4) wICA was applied only to artifact components, then the continuous data were re-constructed into the scalp space, outputting *continuous* data ready for analysis.

Rejected Electrodes Interpolated, Data Segmented into Epochs and Remaining Extremes Rejected:
1) Spherical interpolation of rejected electrodes used to obtain a full set of electrodes for all files
2) Data segmented into epochs for analysis
3) Reject epochs with max-min voltage values >60μV, or kutosis / improbable data >3 overall or >5 at any electrode

Fig. 1. Steps involved in the full RELAX pipeline. Abbreviations: Hz = hertz; s = seconds; PREP = EEG Preprocessing Pipeline; MAD = median absolute deviation; MWF = Multi-channel Wiener filters; ICA = independent component analysis; IQR = inter-quartile range; AMICA adaptive mixture ICA; cudaICA = ICA performed using the cuda cores of a graphic card; fastICA = fast fixed point algorithm ICA; wICA = wavelet enhanced ICA; NaN = not a number.

no further back than the F line, and no more lateral than the 3/4 electrodes). Blinks were marked as the maximum point within each time-period that exceeded the value of the upper quartile of all voltages + the interquartile range (IQR) * 3. A blink artifact mask for MWF cleaning was created by marking the 800 ms surrounding all blink maximums as artifacts.

Third, drift and horizontal eye movements were identified and cleaned with the MWF. For the drift identification step, a copy of the data were re-referenced using PREP's robust average re-referencing approach (Bigdely-Shamlo et al., 2015) as we found drift to be more accurately identified in average re-referenced data. However, the re-referencing was only used at this stage to identify drift, and re-referenced data were not used in the MWF cleaning as our tests showed MWF performed better when implemented prior to average re-referencing. Epochs that showed an amplitude at any electrode > 10MAD from the median of all electrodes were deemed to be affected by drift and marked as artifact periods in the template for MWF cleaning (this approach was adapted from Nolan

et al. (2010), who used the standard deviation (SD) instead of MAD to set the threshold). For detection of horizontal eye movements, a list of the lateral electrodes affected by horizontal eye movements is provided by the user. The RELAX pipeline selected a single electrode from both the left and right side of the head from this list, selecting the first electrode in the list if available, then moving to the second electrode in the list if the first electrode had been rejected, and so on until a horizontal eye movement affected electrode from the list was available. Electrodes used in the current study were "F7″, "FT7″, "F5″, "T7″, "FC5″, "C5″, "TP7″, "AF3″ for the left side, and "F8″,"FT8″,"F6″,"T8″, "FC6″, "C6″, "TP8″, "AF4″ for the right side. While we have provided the user with the ability to set their own electrodes, if the electrodes listed here are available, they reflect good defaults, as the RELAX pipeline with these default settings was extensively tested. If the default electrodes are not available, we recommend prioritizing frontal and maximally lateral electrodes first (such as F7 and F8), and only including more posterior maximally lateral electrodes if these are

not available (such as T7 and T8), followed by less lateral frontal electrodes (such as AF3 and AF4). Periods where the selected electrodes showed voltages > 2MAD from the median of their overall amplitude, with an opposite voltage movement > 2MAD from the median polarity shift in the electrode on the opposite side of the head were assumed to reflect horizontal eye movements (Rogasch et al., 2017). These periods were marked as artifact periods in the MWF template. A delay period of 8 samples was used for all MWF cleaning. This delay period implemented a positive and negative time lag for applying the MWF filter, which turned the spatial MWF into a spatio-temporal finite impulse response filter. RELAX was also set to detect generalised eigenvector deficiency, which sometimes occurs with longer delay periods and can impair the ability of the MWF to clean the data. If generalised eigenvector deficiencies were detected, the algorithm reduced the delay period by a value of 1 sample and ran the MWF again. This approach was repeated up to three times, resulting in a minimum delay period of 5 samples. The full explanation of MWF template construction and cleaning process is provided in the Supplementary Materials (Section 2, pages 6–9).

2.1.3. Wavelet-enhanced independent component analysis

Following the three sequential MWF steps, wICA was used to reduce artifactual ICA components identified by the machine learning algorithm "ICLabel" (Pion-Tonachini et al., 2019), then electrode space data was reconstructed. Firstly, the MWF cleaned data were average re-referenced using PREP's robust re-referencing method, which interpolates missing channels before average re-referencing to prevent reference asymmetries, following the re-removal of the bad channels to prevent rank issues in ICA decompositions (Bigdely-Shamlo et al., 2015). The periods that were marked as extreme outliers (which had previously been marked as NaN for the MWF cleaning) were rejected at this stage. ICA was then computed using either cudaICA (Raimondo et al., 2012), AMICA (Palmer et al., 2012) or fastICA (Hyvarinen, 1999) with the fastICA deflation setting applied to avoid non-convergence issues. ICLabel was used to detect artifactual components (defined by ICLabel as "more likely to reflect any artifact category than to reflect a brain component") (Pion-Tonachini et al., 2019), and only these artifactual components were cleaned with wICA before the continuous data were reconstructed back into the scalp space. The above steps left cleaned and robust average re-referenced continuous data which can then be epoched as desired for different types of analyses.

2.2. Comparison pipelines

To optimize the RELAX pipeline, we tested six potential versions of RELAX with various parameter variations applied. For brevity, we present a summary of the variations of the RELAX pipeline approaches in Table 1. To provide a robust assessment of the performance of our RELAX pipeline, we also compared our pipeline to six of the most commonly reported cleaning approaches. A full explanation of each pipeline is provided in the Supplementary Materials (Section 3, page 12), and in the references cited in the following summary: Firstly, we tested six potential versions of our RELAX pipeline. These included using different ICA algorithms, namely, 1) MWF_wICA_infomax, which used extended-infomax computed in the cuda cores of graphics cards (cudaICA) (Raimondo et al., 2012), 2) MWF_wICA_fastICA, which used the fast fixed point algorithm ICA (fastICA) (Hyvarinen, 1999), or 3) MWF_wICA_AMICA, which used adaptive mixture ICA (AMICA) (Palmer et al., 2012). For brevity, we only report the results of MWF_wICA_infomax in the main manuscript (described simply as MWF_wICA hereafter), since infomax was used as the ICA method for all other pipelines, and the MWF_wICA versions dif-

fered only minimally depending on ICA method. We also tested: 4) MWF_ICA_subtract, which applied MWF cleaning, but then subtracted artifactual ICA components instead of using wICA. Further, we tested: 5) MWF_wICA_CCA, which applied MWF and wICA cleaning, but excluded the cleaning of artifactual components identified as muscle by ICLabel from the wICA step in MWF_wICA, and cleaned muscle activity with the extended canonical correlation analysis (CCA) instead (Janani et al., 2018), and finally: 6) wICA_ICLabel, which excluded MWF cleaning completely and only used wICA applied to artifactual components identified with ICLabel (Pion-Tonachini et al., 2019). This pipeline is similar to those used by Issa and Juhasz (2019) and Mammone et al. (2011), except that our approach identified all artifacts instead of only identifying eye movements (Issa and Juhasz, 2019), and used ICLabel to identify artifacts instead of using entropy and kurtosis measures (Mammone et al., 2011). To help the reader assess the potential of the RELAX pipeline for application to their data, we note here that our recommendation is to use either "**MWF_wICA**" or "**wICA_ICLabel**".

The comparison pipelines we tested were: 7) MWF_only, which was identical to the MWF cleaning steps in our RELAX pipeline but did not apply any additional cleaning after the MWF stage (Somers et al. 2018). 8) MWF_CCA, which applied the sequential MWF cleaning to the data, but instead of applying wICA to this MWF cleaned data as per the RELAX methods, it used the extended CCA to further clean any remaining muscle artifacts (Janani et al., 2018). 9) ICA_subtract, which is probably the most commonly used approach in EEG research – this pipeline computed ICA, then subtracted the components identified as artifacts from the ICA unmixing matrix using ICLabel, then reconstructed the electrode space data (without any MWF cleaning applied) (Pion-Tonachini et al., 2019). 10) The Artifact Subspace Reconstruction (ASR) approach followed by ICA subtraction of artifacts identified by ICLabel (Chang et al., 2019) (referred to as ASR). 11) The Harvard Automated Processing Pipeline for EEG (HAPPE) (Gabard-Durnam et al., 2018). Finally, 12) wICA_all, which cleaned data by applying wICA to all components after the ICA decomposition (Castellanos and Makarov, 2006). For brevity, we present a summary of the comparison pipeline approaches in Table 2, and the full description in the Supplementary Materials.

2.3. Data

We examined the effectiveness of each cleaning pipeline primarily on a large combination dataset (N = 213). The data we used were specifically selected to be challenging to clean, as is often the case for typical EEG data. Data were collected mostly by students, with many data files collected while participants performed cognitive tasks, which produced concentration related forehead and temporal muscle activity, as well as blinks and eye movements. The data also commonly contained bad electrodes, temporarily disconnected electrodes, and other atypical artifacts. As such, the data provided a good "real world" test case for fully automated preprocessing pipelines. In addition, we tested the approach on two additional smaller datasets collected from separate laboratories, with the total data analysed comprising two laboratories, three different EEG systems, and three different tasks (see Supplementary Materials section 5 and 6, pages 54–71 for the results in these additional datasets, as well as our companion article for analysis of a fourth task with RELAX applied to event-related potential analyses: Bailey et al. 2022). All data were recorded from healthy adults with all participants providing informed consent prior to participation. The study was approved by the Ethics Committee of the Alfred Hospital and Monash University. Tasks included the Sternberg task, 2-back task, and Colour-Wheel Recall task (all of which measure

Table 1

A summary of the steps involved in each variation of the RELAX pipeline. Abbreviations: Hz = hertz; s = seconds; PREP = EEG Preprocessing Pipeline; MAD = median absolute deviation; MWF = Multi-channel Wiener filters; ICA = independent component analysis; IQR = inter-quartile range; AMICA adaptive mixture ICA; cudaICA = ICA performed using the cuda cores of a graphic card; wICA = wavelet enhanced ICA.

	MWF_wICA	MWF_ICA_subtract	MWF_wICA_CCA	wICA_ICLabel
Filtering	1 Hz to 80 Hz, with 47–53 Hz notch filter	1 Hz to 80 Hz, with 47–53 Hz notch filter	1 Hz to 80 Hz, with 47–53 Hz notch filter	1 Hz to 80 Hz, with 47–53 Hz notch filter
Bad Channel Rejection	PREP's 'findNoisyChannels', then for each 1 s epoch, reject if > 5 % of data shows extreme values (defined in Fig. 1) or log-power log-frequency slopes > -0.59	PREP's 'findNoisyChannels', then for each 1 s epoch, reject if > 5 % of data shows extreme values (defined in Fig. 1) or log-power log-frequency slopes > -0.59	PREP's 'findNoisyChannels', then for each 1 s epoch, reject if > 5 % of data shows extreme values (defined in Fig. 1) or log-power log-frequency slopes > -0.59	PREP's 'findNoisyChannels', then for each 1 s epoch, reject if > 5 % of data shows extreme values (defined in Fig. 1) or log-power log-frequency slopes > -0.59
Initial Outlying Data Period Rejection	After bad channels are rejected, reject remaining 1 s epochs that exceed the same thresholds as per the bad channel rejection step	After bad channels are rejected, reject remaining 1 s epochs that exceed the same thresholds as per the bad channel rejection step	After bad channels are rejected, reject remaining 1 s epochs that exceed the same thresholds as per the bad channel rejection step	After bad channels are rejected, reject remaining 1 s epochs that exceed the same thresholds as per the bad channel rejection step
Initial Artifact Reduction	3 sequential MWF, cleaning muscle activity first, then eye blinks, then horizontal eye movement and drift	3 sequential MWF, cleaning muscle activity first, then eye blinks, then horizontal eye movement and drift	3 sequential MWF, cleaning muscle activity first, then eye blinks, then horizontal eye movement and drift	None
Second Artifact Reduction	Artifactual ICA components reduced using wICA. ICA computed with fastICA, AMICA, or the cudaICA implementation of the extended infomax ICA (cudaICA results reported in the main manuscript). Artifacts identified by ICLabel.	ICA computed using cudaICA. Artifacts identified by ICLabel, and rejected.	ICA computed using cudaICA. Artifactual ICA components except for muscle activity reduced using wICA (identified by ICLabel), then extended CCA used to clean muscle activity.	ICA computed using cudaICA. Artifactual ICA components identified by ICLabel. Artifacts reduced using wICA.

working memory), and recordings also included both eyes open (EO) and eyes closed (EC) resting-state. For our primary analyses we combined the datasets that used identical recording parameters and would be analysed in similar ways, hence oscillatory analysis approaches were tested using a combined Sternberg and EO/EC resting-state dataset (N = 213). This combination approach was used to provide robust tests of cleaning quality with large datasets.

The combination dataset comprising our primary analyses were recorded using a Neuroscan Synamps2 amplifier with the SCAN 4.3 software interface (Compumedics, Melbourne, Australia) with a 64-channel Quickcap (excluding CB1 and CB2 electrodes) and a sampling rate of 1000 Hz with a 0.01 Hz high-pass and 200 Hz low pass filter. The ground electrode was located at AFz, and the reference was located between Cz and CPz. The first combined dataset we report included EEG recordings from participants completing a Sternberg task (data previously reported by Bailey et al. (2020), and both EO and EC resting-state recordings (total N = 213 after excluding participants who showed very little blink or muscle artifact even prior to cleaning, EO resting-state N = 93, EC resting-state N = 62, data not reported previously, and Sternberg N = 58). Blink metrics were not assessed for the EC resting-state recordings, and 10 files were excluded from blink metric analyses due to not enough blink-locked epochs remaining after the exclusion of epochs that contained multiple blinks (N = 140).

An additional combination of 2-back task related data, as well as EO and EC resting-state recordings, were assessed, reported in the Supplementary Materials (section 5, pages 54–64) (N = 20 healthy adults providing one recording from each type of EEG recording, for a total of 60 EEG files). In this dataset, data were recorded using a Synamps2 amplifier running through the SCAN 4.3 software interface (Compumedics, Melbourne, Australia) with 44 Ag/AgCl electrodes embedded within an EasyCap (Herrsching, Germany). The ground electrode was placed at AFz, and the reference was placed at CPz. A sampling rate of 1000 Hz was used for these recordings, with an online bandpass filter between 0.1 to 200 Hz. Finally, one further dataset of Colour-Wheel Recall task data was tested (N = 23) reported in the Supplementary Materials (section 6, pages 65–71). These data were recorded using a Neu-

roscan amplifier (Compumedics, Melbourne, Australia) with a 62-channel EasyCap (Herrsching, Germany) and a sampling rate of 10 kHz, downsampled to 1000 Hz prior to cleaning. To reduce computation time, MWF_ICA_subtract, MWF_CCA and MWF_wICA_CCA were not tested on these two datasets, as our primary analysis did not indicate these approaches were superior to MWF_wICA.

Various cleaning quality metrics were calculated, using either continuous or epoched data. Continuous data were required to compute Signal-to-Error Ratio (SER) and Artifact-to-Residue Ratio (ARR) values. However, other metrics required epoched data (the total proportion of the epochs rejected by cleaning, and the variance explained by the experimental manipulation). To obtain the epoched data, we interpolated rejected electrodes back into the data (with EEGLAB's 'pop_interp' function and the spherical setting (Delorme and Makeig, 2004)), and applied a typical rejection of epochs based on max–min voltage values > 60 microvolts, or kurtosis / improbable data for all channels > 3 or any channel > 5 to remove any remaining artifacts (since no EEG cleaning pipeline completely addresses all artifacts for all files). This enabled an analysis of differences between the pipelines in the variance explained by differences between commonly analysed experimental conditions (such as between working memory delay and probe periods of the Sternberg task), as well as a measure of the proportion of epochs that were removed by the overall cleaning process, including extreme period rejection and outlier epoch rejection. Less effective cleaning pipelines left more artifacts remaining after cleaning, which were rejected by the final epoch rejection step, providing less data available for analysis after cleaning.

2.4. Cleaning quality evaluation metrics

When assessing the performance of an EEG cleaning method, there are three major challenges. Firstly, the 'ground-truth' of the brain activity signal in isolation is impossible to ascertain in real data, so assessments of cleaning efficacy and signal preservation can only be estimates (Muthukumaraswamy, 2013). Simulated data has been used to address this issue, but it also difficult to ascertain how accurately simulations reflect real data and artifacts,

Table 2

A summary of the steps involved in each of the comparison cleaning pipelines. Abbreviations: Hz = hertz; s = seconds; PREP = EEG Preprocessing Pipeline; MAD = median absolute deviation; MWF = Multi-channel Wiener filters; ICA = independent component analysis; IQR = inter-quartile range; AMICA adaptive mixture ICA; cudaICA = ICA performed using the cuda cores of a graphic card; wICA = wavelet enhanced ICA, ASR = Artifact Subspace Reconstruction; HAPPE = Harvard Automated Processing Pipeline.

	ICA_subtract	MWF_only	MWF_CCA	wICA_all	ASR	HAPPE
Filtering	1 Hz to 80 Hz, with 47–53 Hz notch filter	1 Hz to 80 Hz, with 47–53 Hz notch filter	1 Hz to 80 Hz, with 47–53 Hz notch filter	1 Hz to 80 Hz, with 47–53 Hz notch filter	1 Hz to 80 Hz, with 47–53 Hz notch filter	1 Hz high-pass, CleanLine to remove 50 Hz
Bad Channel Rejection	PREP's 'findNoisyChannels', then for each 1 s epoch, reject if > 5 % of data shows extreme values (defined in Fig. 1) or log-power log-frequency slopes > -0.59	PREP's 'findNoisyChannels', then for each 1 s epoch, reject if > 5 % of data shows extreme values (defined in Fig. 1) or log-power log-frequency slopes > -0.59	PREP's 'findNoisyChannels', then for each 1 s epoch, reject if > 5 % of data shows extreme values (defined in Fig. 1) or log-power log-frequency slopes > -0.59	PREP's 'findNoisyChannels', then for each 1 s epoch, reject if > 5 % of data shows extreme values (defined in Fig. 1) or log-power log-frequency slopes > -0.59	PREP's 'findNoisyChannels', then for each 1 s epoch, reject if > 5 % of data shows extreme values (defined in Fig. 1) or log-power log-frequency slopes > -0.59	Rejects channels with joint probability > 3SD from mean of average log power from 1-125 Hz. Performed twice.
Initial Outlying Data Period Rejection	After bad channels are rejected, reject remaining 1 s epochs that exceed the same thresholds as per the bad channel rejection step	After bad channels are rejected, reject remaining 1 s epochs that exceed the same thresholds as per the bad channel rejection step	After bad channels are rejected, reject remaining 1 s epochs that exceed the same thresholds as per the bad channel rejection step	After bad channels are rejected, reject remaining 1 s epochs that exceed the same thresholds as per the bad channel rejection step	After bad channels are rejected, reject remaining 1 s epochs that exceed the same thresholds as per the bad channel rejection step	None
Initial Artifact Reduction	None	3 sequential MWF, cleaning muscle activity first, then eye blinks, then horizontal eye movement and drift	3 sequential MWF, cleaning muscle activity first, then eye blinks, then horizontal eye movement and drift	None	ASR: automatic detection of clean data periods to determine thresholds, followed by rejection of components with large variance (similar to principal component analysis), then reconstruction of the original electrode space data.	wICA applied to all components
Second Artifact Reduction	Artifactual ICA components subtracted (identified by ICLabel)	None	Extended canonical correlation analysis to clean muscle activity (Janani et al., 2018).	All ICA components reduced using wICA	Artifactual ICA components rejected (identified by ICLabel)	Artifactual ICA components rejected (identified by MARA)

so effective cleaning in simulated data may not translate to effective cleaning in real data (Kumaravel et al., 2022, Mumtaz et al., 2021, Roštáková and Rosipal, 2022). As such, we have limited our analysis to real data, and used a combination of many artifact cleaning metrics and signal preservation metrics. We deliberately selected a complimentary combination of metrics, so some metrics address the potential limitations of other included metrics. We have also included metrics that simply aim to detect whether a well-characterised artifact is still present after cleaning (rather than trying to estimate the artifact signal and how much cleaning reduces that signal). For example, one metric tested whether blink periods are still associated with increased frontal voltage amplitudes after cleaning, and another whether epochs still show log-power log-frequency slopes that indicate muscle activity remains after cleaning. This approach avoids the 'ground-truth' issue by simply asking "are the well-known artifact characteristics still present in the data after cleaning?". Secondly, EEG data is contaminated by many different artifact types, all of which should be cleaned effectively for optimal data analysis. To address this, we have included artifact reduction metrics that address all common artifacts. Thirdly, perhaps the most important metrics should assess the impact of EEG data cleaning on the practical outcomes of the research (Clayson et al., 2021a). As such, we have included multiple metrics that assess whether cleaning by the different pipelines leads to more variance explained by different experimental manipulations.

To ensure we assessed the different cleaning pipelines fully for effectiveness at cleaning both the full range of potential artifacts, and for preserving the neural signal (not over-cleaning the data),

we tested the pipelines across six different types of cleaning quality metrics, for a total of 13 different metrics. All metrics are explained in detail in the Supplementary Materials (section 3, pages 13–18) and summarised in Table 3. Briefly, these metrics included: the SER: a measure of the amount of signal left unaffected during the clean EEG periods after cleaning. The SER is obtained by dividing the expected value operator of the squared signal amplitude in periods marked as clean in each electrode in the raw EEG by the squared signal of the signal that was removed by cleaning during the periods marked as clean the MWF templates (Somers et al., 2018). The individual electrode values are then averaged across electrodes with weighting applied by the artifact signal in each electrode proportional to the artifact signal in all electrodes, so electrodes with larger artifact contribute more to the SER measure (Somers et al., 2018). Secondly, we included a measure of the extent to which all artifacts identified by our MWF template were reduced – the ARR. The ARR was calculated by obtaining the expected value operator of the square of the removed artifact, divided by the expected value operator of the square of the raw data from the artifact periods minus the removed artifact signal from the artifact periods (Somers et al., 2018). Effective cleaning reflects reductions in large artifact signals, such that the denominator is small and high ARR values result from the division. As such, high values for the SER and ARR indicate good performance (Bertrand, 2015, Somers and Bertrand, 2016, Somers et al., 2018). It is important to note that SER and ARR are complementary metrics. To achieve the highest cleaning quality a pipeline should produce both high SER and high ARR values; it would be easy to obtain very high ARR values and very low SER values by being

excessively stringent in artifact reduction and not concerned with preserving the neural signal, but this would reduce the neural activity of interest (Somers et al., 2018).

Additionally, we measured the ratio of blink amplitude to non-blink periods after cleaning, referred to as the Blink Amplitude Ratio (BAR) (Robbins et al., 2020). We examined BAR averaged across the frontal electrodes affected by blinks (fBAR), and across all electrodes (allBAR, reported in the Supplementary Materials, section 4 page 25). For fBAR and allBAR, values near to 1 indicate optimal performance, with activity that is time-locked to blinks showing the same amplitude as non-blink activity after cleaning, while values < 1 reflect overcleaning, and values > 1 reflect under cleaning (Robbins et al., 2020). Next, we measured the number of epochs showing log-power log-frequency slopes > -0.59 that indicated the epoch was still contaminated by muscle activity (Fitzgibbon et al., 2016), and the severity by which these slopes exceeded the muscle threshold within epochs showing muscle activity remaining after cleaning (these results are reported in the Supplementary Materials, section 4, page 30). Higher values reflect poorer performance for these metrics. The final cleaning metric we assessed was the proportion of epochs that were rejected through the cleaning process against the initial number of epochs in the raw data. While this measure can be affected by the raw data quality (with lower raw data quality requiring more data rejection), comparisons of the number of epochs rejected by different cleaning pipelines (applied to the same datasets) indicate which pipelines are better at cleaning data (such that fewer bad epochs need to be removed, and more epochs are available for analysis). Providing higher quality data with more epochs for analysis can be more important for statistical power than recruiting more participants (Kolossa and Kopp, 2018). As such, lower values for the proportion of epochs rejected reflect better performance.

Next, and perhaps most importantly, we assessed the amount of variance explained by a variety of experimental manipulations after cleaning by each pipeline. We chose experimental manipulations that are well established to provide differentiation of neural activity in the comparison of two conditions, which assesses the real-world applicability of the cleaning pipelines (Clayson et al., 2021a). We assessed the variance explained by a comparison of alpha oscillatory power between the working memory delay and probe periods of the Sternberg task (Bailey et al., 2020) and alpha power between EO and EC resting-state data (computation of which is described in the Supplementary Materials, section 4, pages 36–53). Within these metrics, higher values are likely to reflect better performance, with cleaning pipelines that provided higher values providing a better chance to detect statistically significant differences within an experiment.

2.5. Statistics

In order to compare the effectiveness of artifact cleaning between the cleaning pipelines across each of the metrics, we used the robust repeated measures ANOVA function "rmanova" from the WRS2 package in R (R Core Team v4.0.4) (Mair and Wilcox, 2020). These tests are robust against the normality and homoscedasticity assumptions of traditional parametric statistical tests, while still providing equivalent power (Mair and Wilcox, 2020). When the overall ANOVA was significant, pairwise comparisons of the cleaning pipelines were performed using the robust post-hoc t-test function "rmmcp". This applies multiple comparison controls across the post-hoc tests within each omnibus ANOVA using Hochberg's approach to control for the family-wise error (Mair and Wilcox, 2020). However, experiment-wise multiple comparison controls were not implemented as we deemed it was more important to provide sensitivity to differences in cleaning outcomes than to protect against false positive results when

exploring a wide range of EEG cleaning methods for the optimal approach (Bender and Lange, 2001). Additionally, we tested a large number of metrics to demonstrate which pipelines provide the best overall performance (rather than for a single artifact type). We also used multiple datasets to provide an indication of how replicable our results are across datasets. Experiment-wise multiple comparison controls in this large number of comparisons across this wide range of metrics and across multiple datasets would severely reduce the sensitivity of our study to detect important differences between cleaning approaches. It would have also adversely encouraged us to reduce the number of metrics we assessed, leading to a less well tested final pipeline. As such, we emphasised sensitivity to detect meaningful differences between pipelines over the potential for false positives. To visualise the cleaning efficacy data, we provide raincloud plots with boxplot medians and IQR marked for the reader to see the full data with minimal distortion (Allen et al., 2019). This includes outliers, which are important to note, as outliers indicate individual files in which a cleaning method might not have adequately cleaned the artifact.

In addition to the robust statistics used to compare artifact cleaning metrics, we have provided graphs of the amount of variance explained (np^2) by the comparison between the two experimental conditions for each pipeline for the comparisons of the amount of variance explained by the experimental manipulation (as well as the absolute values for each condition from each pipeline in the Supplementary Materials, section 4, pages 36–54). To test for differences between pipelines in np^2 for the experimental manipulation, we tested the interaction between each pair of pipelines and the two experimental conditions using the Randomisation Graphical User Interface (RAGU) Global Field Potential (GFP) test which measures overall neural response amplitudes (Habermann et al., 2018, Koenig et al., 2011). We also used the Topographical Analysis of Variance (TANOVA) with L2 norm tests, which measures the distribution of neural activity after normalisation for amplitude (Habermann et al., 2018, Koenig et al., 2011). We have provided heat maps displaying np^2 values for each interaction in the Supplementary Materials (section 4, pages 36–55), with the multiple comparison controls applied across the post-hoc tests within each omnibus ANOVA for each measure using the false discovery rate (FDR-p) (Benjamini and Hochberg, 1995).

We also provide a rank order of the means from each pipeline for each cleaning efficacy and variance explained metric, with significant differences noted (Table 4). We have also provided means and SD tables, and specific details of post-hoc tests within a heat map, including confidence intervals for t-test comparisons with significant tests highlighted for all metrics in the Supplementary Materials (section 4). Finally, we have provided scatter plots of mean SER × ARR values for each pipeline across the different datasets, so that these two metrics can be considered together for each cleaning pipeline.

3. Results

All metrics we tested showed a significant difference in the omnibus ANOVA (all p < 0.01). A rank order of the means from each pipeline (with significant differences as detected by the post-hoc t-tests noted) can be viewed in Table 4. Due to the high number of comparisons, for readability, we provide here a summary of only the results that were most relevant to the evaluation of the pipelines for cleaning efficacy and the detection of differences between well-established experimental conditions. Full details of our primary analyses are provided in the Supplementary Materials (section 4, pages 20–53, Figures S1-S34). Additionally, the details of the analyses of the non-primary datasets (the combined 2back, EO and EC dataset and the Colour Wheel task dataset) are reported

Table 3

Description of cleaning quality evaluation metrics used for comparing pipelines. Not all results from the pipelines are included in the main manuscript but can be viewed in the Supplementary Materials. Metrics denoted with an asterisk are only presented in the Supplementary Materials. RMS: Root Mean Square, TANOVA: topographical analysis of variance.

	Metric Name	Summary of Metric Purpose	Required EEG Data Type	Evaluation of Metric
The following metrics were implemented to assess the effectiveness of each pipeline at cleaning a range of potential EEG artifacts.				
Data Cleaning Metrics	**Signal to Error Ratio (SER)**	Indicates how well the signal in clean EEG periods preserved by the cleaning process	Continuous	Higher values reflect better cleaning
	Artifact to Residue Ratio (ARR)	Indicates the extent to which all artifacts were reduced	Continuous	Higher values reflect better cleaning
	Frontal Electrodes Blink Artifact Ratio (fBAR)	Indicates whether cleaning has effectively cleaned the blink artifact in frontal electrodes (with the ability to indicate over-cleaning).	Continuous	Values $\sim$=1 reflect optimal cleaning, values < 1 indicate overcleaning, and values > 1 reflect under cleaning.
	All Electrodes Blink Artifact Ratio (allBAR)*	Indicates whether cleaning has effectively cleaned the blink artifact in all electrodes (with the ability to indicate over-cleaning).	Continuous	Values $\sim$=1 reflect optimal cleaning, values < 1 indicate overcleaning, and values > 1 reflect under cleaning.
	Muscle Activity Remaining After Cleaning	Indicates the proportion of the data containing likely muscle activity after cleaning	Continuous	Higher values reflect poorer performance in removing muscle activity
	Proportion of Epochs Removed by Cleaning	Reflects the proportion of epochs rejected by the overall cleaning process.	Epoched	Higher values reflect a larger proportion of epochs rejected by the cleaning pipeline.
The following metrics were implemented to depict the ability of the different pipelines to differentiate experimental conditions by examining the interactions between different pairs of pipelines and the condition of interest.				
Variance Explained by between condition comparisons of Alpha Oscillation Power	**Resting data Root Mean Squared (RMS) test**	Reflects the variance in alpha power overall neural response strength explained by the difference between eyes-open and eyes-closed resting	Epoched	Higher values reflect better performance
	Resting data Topographical Analysis of Variance (TANOVA) test	Reflects the variance in alpha power distribution explained by the difference between eyes-open and eyes-closed resting in alpha power TANOVA (normalised within each participant for overall amplitude)	Epoched	Higher values reflect better performance
	Working Memory delay vs probe periods RMS 250 to 150 ms	Reflects the variance in alpha power overall neural response strength explained by the difference between delay and probe periods in the Sternberg task	Epoched	Higher values reflect better performance
	Working Memory delay vs probe periods TANOVA 0 to 750 ms	Reflects the variance in alpha power distribution explained by the difference between delay and probe periods in the Sternberg task immediately after stimuli presentation	Epoched	Higher values reflect better performance
	Working Memory delay vs probe periods TANOVA 750 to 2000 ms	Reflects the variance in alpha power distribution explained by the difference between delay and probe periods in the Sternberg task from 750 m to 2000 ms after stimuli presentation	Epoched	Higher values reflect better performance

in the Supplementary Materials (sections 4–5, pages 54–71, Figures S35-S60). Examples of the raw EEG trace and the EEG trace after cleaning with each of the pipelines are also presented in the Supplementary Materials (Figures S61-S71). In addition to the results reported below, we also note that RELAX is quick to run - performing MWF_wICA using fastICA without computing cleaning quality metrics took 225 seconds, on a 60 electrode 7-minute EEG file, recorded at 1000 Hz using an Intel Core i7-10875H CPU @ 2.3 GHz, 32 GB RAM. Performing wICA_ICLabel without metrics only took 140 seconds. Computing cleaning quality metrics only added 25 seconds to the computation time.

3.1. Artifact cleaning metrics

3.1.1. Signal-to-Error Ratio and Artifact-to-Residue Ratio

SER and ARR values can be viewed in Fig. 2 (and Figures S1-S5 in the Supplementary Materials). When SER and ARR values were viewed together (Fig. 3), it was apparent that MWF_wICA and MWF_wICA_CCA performed better than ASR in both metrics. MWF_wICA and MWF_wICA_CCA also performed equally to ICA_subtract in SER, while at the same time performing better in the ARR metric. MWF_ICA_subtract provided higher ARR values but at the expense of lower SER values than MWF_wICA. MWF_ICA_subtract also provided similar SER and higher ARR values com-

pared to ASR suggesting that MWF cleaning is superior to ASR as an initial data cleaning step prior to using ICA to subtract artifacts. HAPPE and wICA_all showed very high ARR values, but at the expense of very low SER values. Inversely, MWF_only, MWF_CCA, and wICA_ICLabel showed high SER values but lower ARR values (MWF_only performed the best of these three pipelines, providing high SER and a similar ARR value to MWF_CCA and wICA_ICLabel). Finally, note that wICA_ICLabel outperformed ICA_subtract when viewing SER and ARR concurrently. While these initial results might suggest MWF_only is the best pipeline, the SER and ARR values should be viewed in the context of the other metrics we have included, which indicate how much blink and muscle artifacts still affect the data after cleaning, as well as the variance explained by the experimental manipulations.

3.1.2. Blink amplitude ratios

MWF_wICA and MWF_wICA_CCA performed significantly better than all other pipelines (while not differing from each other), providing values closest to 1 for both fBAR (Fig. 4, and Figures S6-S7 in the Supplementary Materials) and allBAR (reported in the Supplementary Materials, section 3, page 25, Figures S8-S9). This was with the exception of MWF_ICA_subtract and MWF_CCA, which only performed significantly worse than MWF_wICA and MWF_wICA_CCA for allBAR values (and did not significantly differ

Table 4

Rank order from best performance (on the left) to worst performance (on the right) (by mean). Note that for some metrics, better performance is reflected by lower values, for example the proportion of epochs showing muscle activity remaining after cleaning, and these metrics are ordered from best performance (on the right) to worst performance (on the left) rather than from larger to smaller values. Significant differences are highlighted for which pipelines performed significantly better than other pipelines using the following notation for ease of understanding: better performance > worse performance (rather than higher values > lower values). Because sometimes pipeline 1 differed from pipeline 2, but pipeline 3 did not differ from either 1 or 2, we have used the following notation: ^ = significantly higher than the pipeline marked with a ^^ within the same section (while the others in the category are not significantly different from each other). * = significantly higher than the pipeline marked with a ** in the same category, and so on for the following symbols: $^{+@\$!^*}$. Please refer to Tables 1 and 2 for an explanation of the labels of the EEG cleaning pipelines presented in this figure. Note that MWF_wICA_fastICA and MWF_wICA_AMICA refer to the implementations of MWF_wICA that use the fastICA and AMICA ICA algorithms respectively.

Oscillation Dataset Cleaning Metrics	
Signal to Error Ratio	MWF_only > MWF_CCA > wICA_ICLabel > MWF_wICA_AMICA > MWF_wICA_fastICA*, MWF_wICA, ICA_subtract, MWF_wICA_CCA** > ASR, MWF_ICA_subtract > HAPPE > wICA_all
Artifact to Residue Ratio	HAPPE > wICA_all > MWF_ICA_subtract > MWF_wICA_fastICA, MWF_wICA, MWF_wICA_CCA > MWF_wICA_AMICA > ASR > MWF_CCA > wICA_ICLabel > MWF_only, ICA_subtract
Frontal Blink Amplitude Ratio	MWF_wICA_fastICA, MWF_wICA, MWF_wICA_CCA, MWF_ICA_subtract, MWF_wICA_AMICA > MWF_CCA^{@+}, wICA_all*, ASR^{@}, ICA_subtract^{@@}, HAPPE^{***}, MWF_only^{++}, wICA_ICLabel^^{@@}
Proportion of epochs showing muscle slopes after cleaning	MWF_ICA_subtract > MWF_wICA > MWF_wICA_fastICA > MWF_wICA_AMICA > MWF_wICA_CCA > MWF_CCA > ICA_subtract > ASR, wICA_ICLabel > MWF_only > HAPPE > wICA_all
Proportion of Epochs Removed by Cleaning	HAPPE > wICA_all > MWF_wICA_fastICA^, MWF_wICA *, MWF_ICA_subtract*^{+^^}, MWF_wICA_AMICA*, MWF_wICA_CCA^^{*^{*@}}, MWF_CCA^^{*^{+**}}, MWF_only^^{*^{+**@@}}, ICA_subtract^^{*^{+**}}, wICA_ICLabel^^{*^{+**}}. ASR showed the highest mean value, but only significantly differed from HAPPE due to the large variance in proportion of epochs removed for the ASR pipeline.
Variance explained in Alpha Oscillation Power Comparisons	
Resting data Root Mean Squared (RMS) Test	ASR*, MWF_wICA_fastICA*^{@}, MWF_wICA *^{@}, MWF_wICA_AMICA*^{@}, MWF_ICA_subtract*^{+@@}, MWF_wICA_CCA*, MWF_only^{+++}, MWF_CCA^{+++}, HAPPE**^, wICA_ICLabel^^, ICA_subtract^^ > wICA_all
Resting data Topographical Analysis of Variance (TANOVA) Test	wICA_all > MWF_CCA > MWF_only > MWF_wICA *, MWF_wICA_AMICA*^, MWF_wICA_fastICA*^{+}, MWF_wICA_CCA*^{+}, MWF_ICA_subtract*^^, wICA_ICLabel**^^, ASR, ICA_subtract**. HAPPE provided the lowest variance explained, but significantly differed from all pipelines *except* MWF_CCA and MWF_only (which showed the 2nd and 3rd largest np² values). This odd result may be due to the multidimensional nature of interactions between pipeline and EO/EC conditions in the TANOVA, combined with the fact the TANOVA compares the distribution of activity (rather than a single value). As such, the result may be due to a more similar match in topographical difference between EO and EC conditions for HAPPE, MWF_CCA, and MWF_only compared to the other pipelines.
Working Memory delay vs probe periods Root Mean Squared Test from 250 to 1500 ms	HAPPE > wICA_all > ASR^{+}, ICA_subtract*, wICA_ICLabel*, MWF_wICA ^{+***^}, MWF_ICA_subtract^{@****}, MWF_only^{!@@**^^}, MWF_CCA^{!@@**+^^}, MWF_wICA_AMICA ^{****}, MWF_wICA_CCA^{@@**}, MWF_wICA_fastICA^{+***!!}
Working Memory delay vs probe periods Topographical Analysis of Variance Test from 0 to 750 ms	HAPPE*, wICA_ICLabel^, ICA_subtract^, wICA_all^{+^^}, MWF_only^{@**^^++}, MWF_CCA^{@**^^++}, ASR^{!**^^+@@}, MWF_wICA_AMICA^{!**^^++@@}, MWF_wICA_fastICA**^{^^+@@!!}, MWF_wICA_CCA**^{^^+@@}, MWF_wICA **^{^^+@@}, MWF_ICA_subtract**^{^^+@@}
Working Memory delay vs probe periods Topographical Analysis of Variance Test from 750 to 2000 ms	ICA_subtract, wICA_ICLabel > wICA_all > MWF_only > MWF_CCA > MWF_wICA_CCA, MWF_wICA, MWF_wICA_fastICA, MWF_wICA_AMICA, MWF_ICA_subtract > HAPPE

in fBAR values). HAPPE, MWF_only, and wICA_ICLabel were the worst performers for fBAR, and MWF_only, ASR, ICA_subtract and wICA_ICLabel were the worst performers for all_BAR. However, all pipelines performed reasonably well with mean fBAR and all-BAR values ranging from 1.015 to 1.077. Having said this, it is important to note the presence of outliers, which indicated where a pipeline did not clean blinks effectively for a specific EEG file. Visual inspection of the raincloud plot indicated MWF_wICA and MWF_ICA_subtract showed the fewest outliers.

3.1.3. Muscle activity remaining after cleaning

MWF_ICA_subtract outperformed all other pipelines in terms of the proportion of epochs showing muscle activity remaining after cleaning (essentially 0 for almost all EEG files, Fig. 5, and Figures S10-S12 in the Supplementary Materials). MWF_wICA performed the next best (with values of almost 0 for most EEG files, but some outlying files showed more epochs contaminated by muscle activity), followed by MWF_wICA_CCA, then MWF_CCA then ICA_subtract. ASR and wICA_ICLabel performed worse, followed by MWF_only. This overall rank order was similar for the severity of muscle slopes that exceeded the threshold after cleaning, with MWF_ICA_subtract performing the best, followed by MWF_wICA (reported in the Supplementary Materials, section 4, page 30, Figures S13-S14). Of note, both the HAPPE and wICA_all pipelines showed considerably more epochs identified as containing residual muscle activity after cleaning than other pipelines (>75 % of epochs

compared to < 10 % for other pipelines). Instead of reflecting genuine residual muscle activity, this may reflect an overall flattening of the spectra slope due to considerable removal of power in the lower frequencies by these pipelines (producing log-power log-frequency slope values more like muscle affected data). However, these pipelines showed relatively more beta range power in the power-frequency plot, a feature which is not present in the other pipelines.

3.1.4. Proportion of epochs removed by all cleaning steps

HAPPE and wICA_all produced data that had the fewest epochs removed by cleaning (Fig. 6, and Figures S17-S18 in the Supplementary Materials). These pipelines also produced very low amplitude cleaned data, so we suspect their high performance in this metric is the result of producing cleaned data with no epochs that exceeded the epoch rejection criteria in our final epoch rejection cleaning step. MWF_wICA then MWF_ICA_subtract were the next best performers, followed by MWF_wICA_CCA, MWF_CCA, MWF_only, ICA_subtract and wICA_ICLabel. ASR showed the highest mean value (worst performance), but only significantly differed from HAPPE (we suspect this was due to its high variance which can be seen in Fig. 6). When we tested for post-hoc differences between individual pipelines using a bootstrap pairwise t-test (pairdepb in the WRS2 package), ASR showed a higher proportion of epochs rejected than all other pipelines.
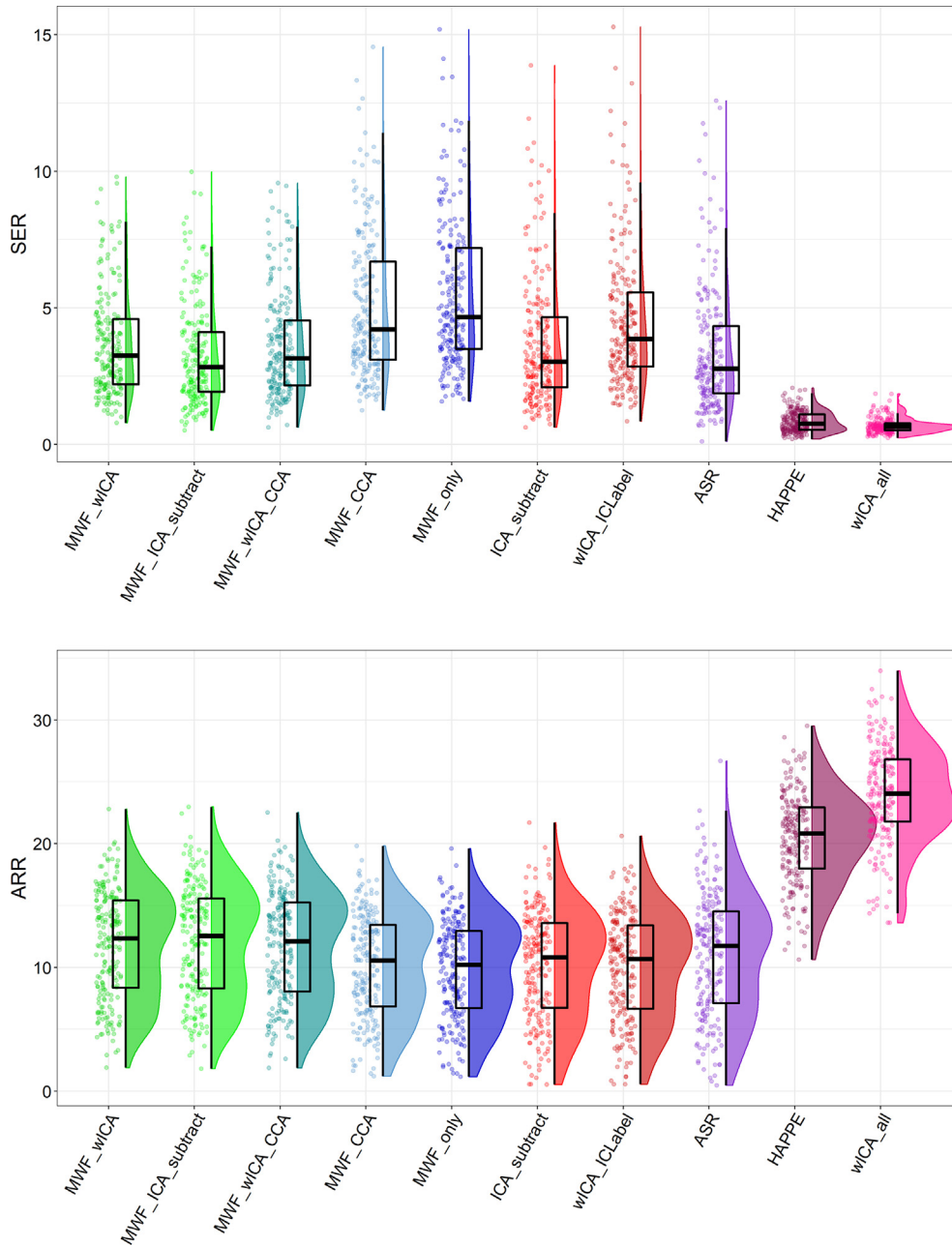
Fig. 2. Raincloud plots depicting Signal-to-Error (SER) and Artifact-to-Residue (ARR) values from the combined eyes open, eyes closed, and Sternberg data (N = 213) for each of the cleaning pipelines. Please refer to Tables 1 and 2 for an explanation of the labels of the EEG cleaning pipelines presented in this figure. Higher values in both metrics concurrently reflects better cleaning (high SER reflects good signal preservation, high ARR reflects good artifact reduction).

3.2. Variance explained by experimental manipulations

3.2.1. Variance explained by the difference between Eyes-Open and Eyes-Closed Resting-State

Fig. 7 depicts Fourier frequency-power spectrum plots for EO and EC resting-state, and Fig. 8 depicts the amount of variance explained by the difference between EO and EC resting-state in root mean squared (RMS; reflecting overall neural response strength) alpha power and TANOVA tests (reflecting the distribution of neural activity) across the different cleaning pipelines (with the full details of each comparison provided in Figures S19-S24 in the Supplementary Materials). When comparing the difference in EEG signals between resting-state conditions, all pipelines showed the expected pattern of higher alpha-band power during EC com-

pared to EO. For measures of the variance explained by the difference between EO and EC resting-states in the global strength of alpha power across all electrodes (RMS), ASR showed the highest np^2 value, but did not significantly differ from MWF_wICA, MWF_ICA_subtract, MWF_wICA_CCA, wICA_ICLabel, or ICA_subtract. MWF_wICA also outperformed MWF_ICA_subtract, HAPPE and wICA_all.

With regards to the distribution of alpha activity (TANOVA), wICA_all showed the highest np^2 value, followed by MWF_CCA, MWF_only then MWF_wICA. ASR, wICA_ICLabel, ICA_subtract, and HAPPE showed the lowest np^2 values, with the later three of these pipelines showing significant differences compared to all other pipelines, with the unusual exception of HAPPE, which significantly differed from all pipelines except MWF_CCA and
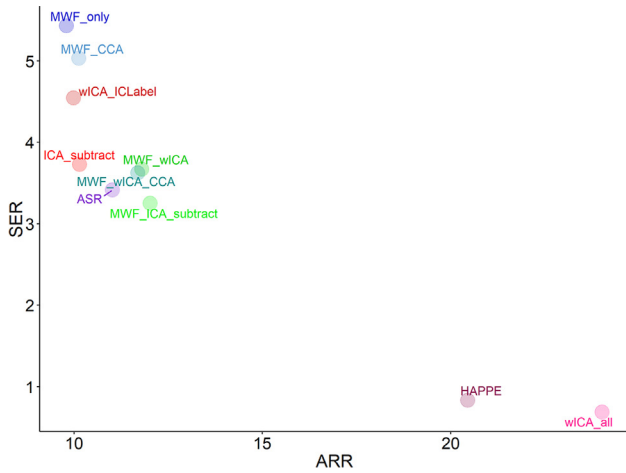
Fig. 3. A scatterplot depicting both mean Signal-to-Error (SER) and mean Artifact-to-Residue (ARR) values for the resting-state eyes open, eyes closed, and Sternberg dataset from each cleaning pipeline. Please refer to Tables 1 and 2 for an explanation of the labels of the EEG cleaning pipelines presented in this figure. Higher values in both metrics concurrently reflects better cleaning (high SER reflects good signal preservation, high ARR reflects good artifact reduction).

MWF_only (pipelines that showed the 2nd and 3rd largest np^2 values). Interpreting interactions in the normalised distribution of activity is complicated, but we suspect this finding might be due to a more similar pattern of differences in the distribution of activity between the EO and EC activity for HAPPE, MWF_CCA and MWF_only (while other pipelines may have showed a different pattern to HAPPE, and more variance explained than HAPPE but less than MWF_only and MWF_CCA). The distribution of activity from each pipeline can be viewed in the Supplementary Materials (Figures S22-23, pages 41–42).

3.2.2. Variance explained by the experimental manipulation in the Sternberg task

Fig. 9 depicts the Morlet wavelet transform power-frequency spectrum plots for the Sternberg delay and probe periods. Note that while the values produced by Fourier frequency-power are μV^2, because Morlet wavelet transform power is dependent on the settings used, Morlet wavelet transform power values are defined in "arbitrary units" (a.u.), and the power values produced by these two transforms are on entirely different scales and are not comparable. Fig. 10 depicts the amount of variance explained by the difference between the Sternberg delay and probe periods in the Morlet wavelet transform alpha power RMS test across the different cleaning pipelines (which compared overall neural response strength from 250-1500 ms after the stimuli) and the TANOVA from 0 to 750 ms and 750 to 2000 ms after the stimuli (with the full details of each comparison provided in Figures S25-S34 in the Supplementary Materials). All pipelines showed higher alpha power during the Sternberg delay period than the probe period, as expected. With regards to the alpha power RMS, HAPPE showed the highest np^2 value for the difference between the two conditions, followed by wICA_all. ASR, ICA_subtract, and wICA_ICLabel showed the next highest, significantly larger than MWF_wICA, MWF_ICA_subtract, MWF_only, MWF_CCA, and MWF_wICA_CCA. While the HAPPE and wICA methods provided the largest np^2 values, these two pipelines resulted in alpha power values > 2 orders of magnitude lower than the other pipelines (with a similar reduction in power across all other frequencies). This aligns with HAPPE and wICA_all showing the lowest SER values, suggesting much of the neural activity signal was eliminated by these methods. wICA_all and HAPPE also produced a different topography of alpha activity to the other pipelines, with smaller differences in alpha power between Sternberg retention and probe conditions at PO7/8 electrodes, where differences are usually the largest (see Figs. 9, 11 and 12 and Supplementary Materials Figures S31-32, pages 41–52). As such, we suspect the application of wICA to all components in these methods may remove prominent characteristics of the EEG activity, so that although these methods might provide more power to detect differences between experi-
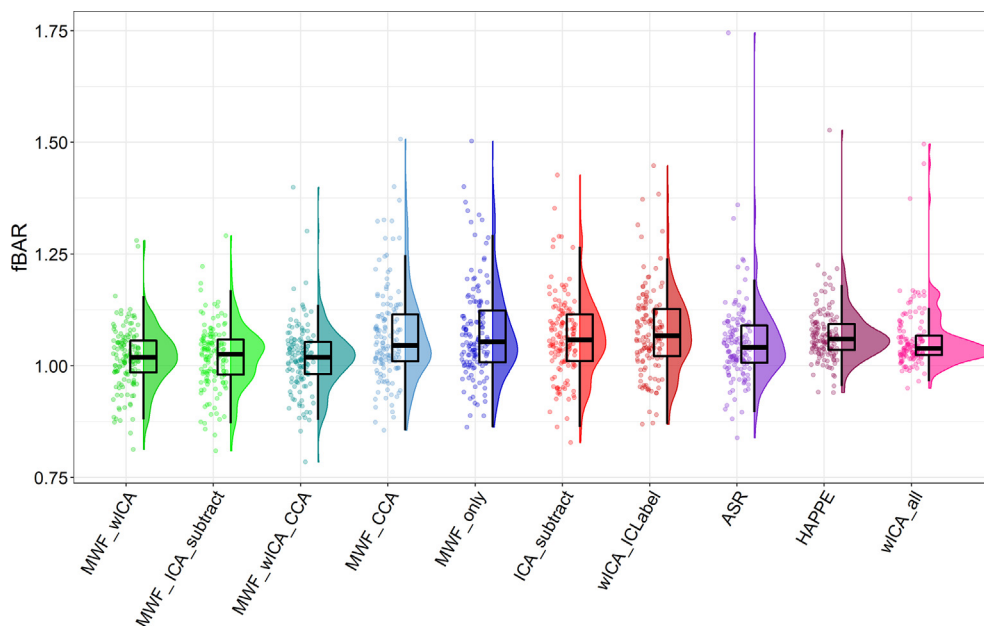

Fig. 4. Raincloud plot depicting frontal blink amplitude ratio (fBAR) values from the combined eyes open, and Sternberg data (N = 140) for each of the cleaning pipelines. Please refer to Tables 1 and 2 for an explanation of the labels of the EEG cleaning pipelines presented in this figure. Values of 1 reflect blink cleaning so that blink periods are no longer larger in amplitude than the surrounding data. Values < 1 are likely to reflect overcleaning, and values > 1 are likely to reflect blink artifact remaining after cleaning.
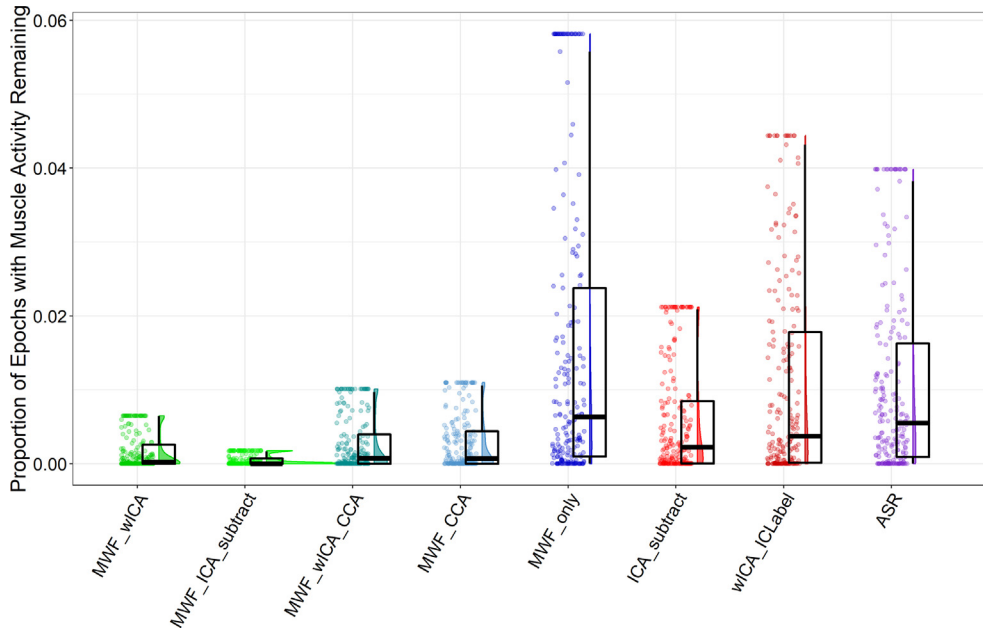
Fig. 5. Raincloud plot depicting the proportion of epochs showing log-power log-frequency values above the −0.59 threshold from the combined eyes open, eyes closed, and Sternberg data (N = 213) for each of the cleaning pipelines. Note that this figure excludes HAPPE and wICA_all, as these pipelines showed median values > 0.75 and made the scale of the graph such that it was difficult to visualise differences in the other pipelines. Note also that we have winsorized the data in the figure, as the outliers also made the scale such that it was difficult to visualise differences in the other pipelines. The full data can be viewed in the Supplementary Materials Figure S11 (page 30). Please refer to Tables 1 and 2 for an explanation of the labels of the EEG cleaning pipelines presented in this figure. Lower values reflect more effective cleaning of muscle activity, with values of 0 indicating no epochs containing muscle artifact remain after cleaning.

Fig. 6. Raincloud plot depicting the proportion of epochs in the data rejected from the combined eyes open, eyes closed, and Sternberg data (N = 213) for each of the cleaning pipelines. Please refer to Tables 1 and 2 for an explanation of the labels of the EEG cleaning pipelines presented in this figure. Lower values indicate that the cleaning pipeline preserves more epochs of the data.

mental conditions in some circumstances, they may do so by only providing a partial characterization of the neural activity recorded to the EEG data. However, despite the reduction in signal, both HAPPE and wICA cleaned data still showed the typical pattern of activity as the peak frequency, so if alpha power were computed relative to the other frequencies, HAPPE and wICA may still show similar results to the other pipelines (although the prominence of

alpha above other frequencies was reduced in HAPPE and wICA compared to other pipelines).

With regards to comparisons of the distribution of activity (assessed by the TANOVA), two separate time periods showed a difference in the distribution of alpha activity between the Sternberg delay and probe period in all pipelines – first a large difference from 0 to 750 ms, then a smaller difference from 750 to 2000 ms. For the first time-period (0 to 750 ms), HAPPE, wICA_ICLabel and
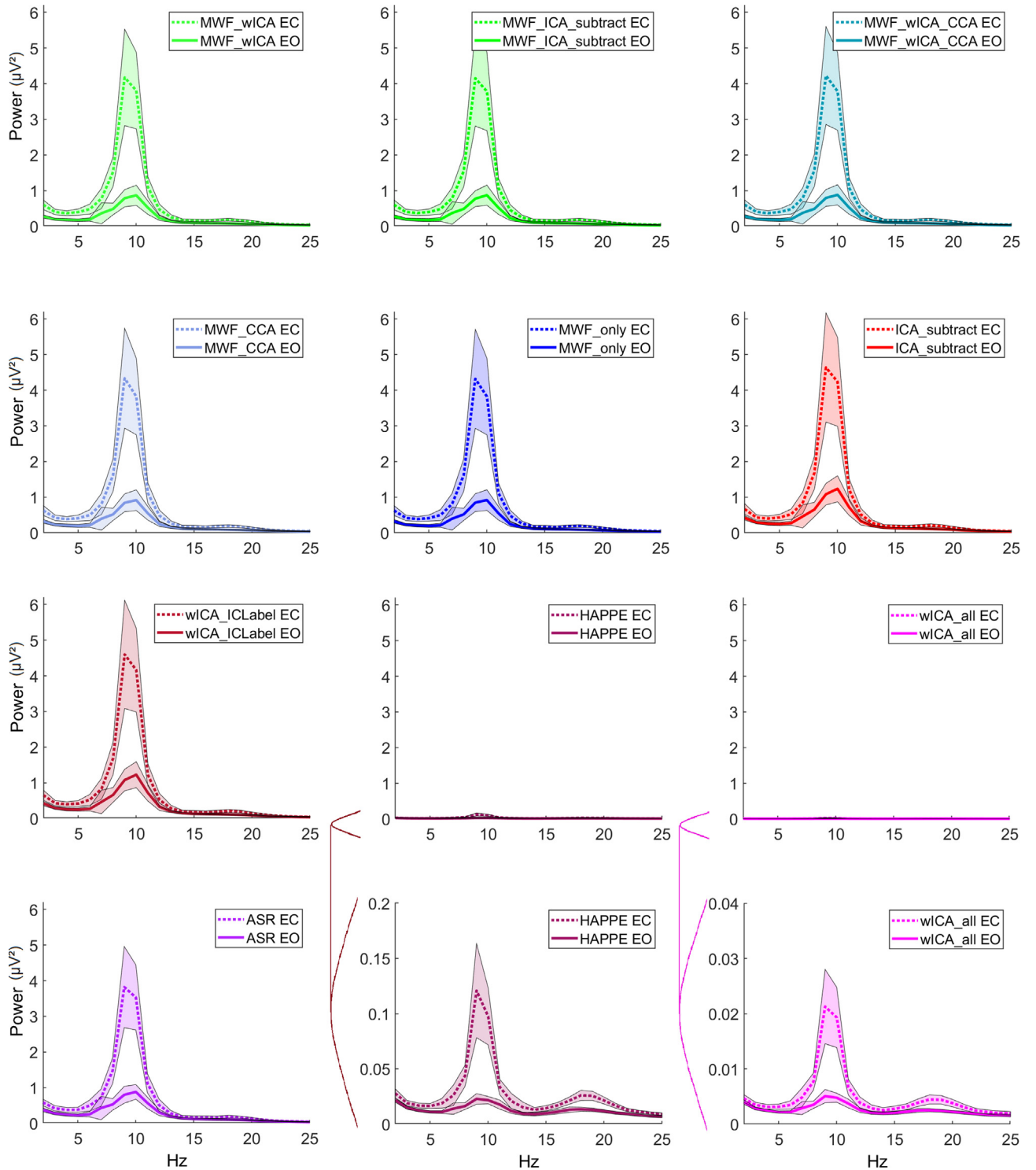
Fig. 7. Frequency power plots depicting fast-Fourier transformed power averaged across PO7 and PO8 at each frequency from the eyes open (EO) and eyes closed (EC) resting-state EEG recordings for each cleaning pipeline (shaded errors reflect 95% confidence intervals). Note the different scale required to see the power spectrum for HAPPE and wICA_all. Please refer to Tables 1 and 2 for an explanation of the labels of the EEG cleaning pipelines presented in this figure.

ICA_subtract showed the best performance, significantly better than wICA_all, which outperformed MWF_only, and MWF_CCA, which outperformed ASR, MWF_wICA, MWF_wICA_CCA, and MWF_ICA_subtract. However, it is worth noting that for this metric, all pipelines produced np^2 values between 0.24 and 0.28, so the differences were not large. For the second time-period examined with the TANOVA test (750 to 2000 ms), a different pattern was

apparent, with ICA_subtract and wICA_ICLabel performing the best, followed by wICA_all, which performed better than MWF_only, MWF_CCA, MWF_wICA_CCA, MWF_wICA, and MWF_ICA_subtract. HAPPE showed the worst performance. However, similar to the earlier time-period, all pipelines produced np^2 values between 0.18 and 0.25, so the differences were not large.
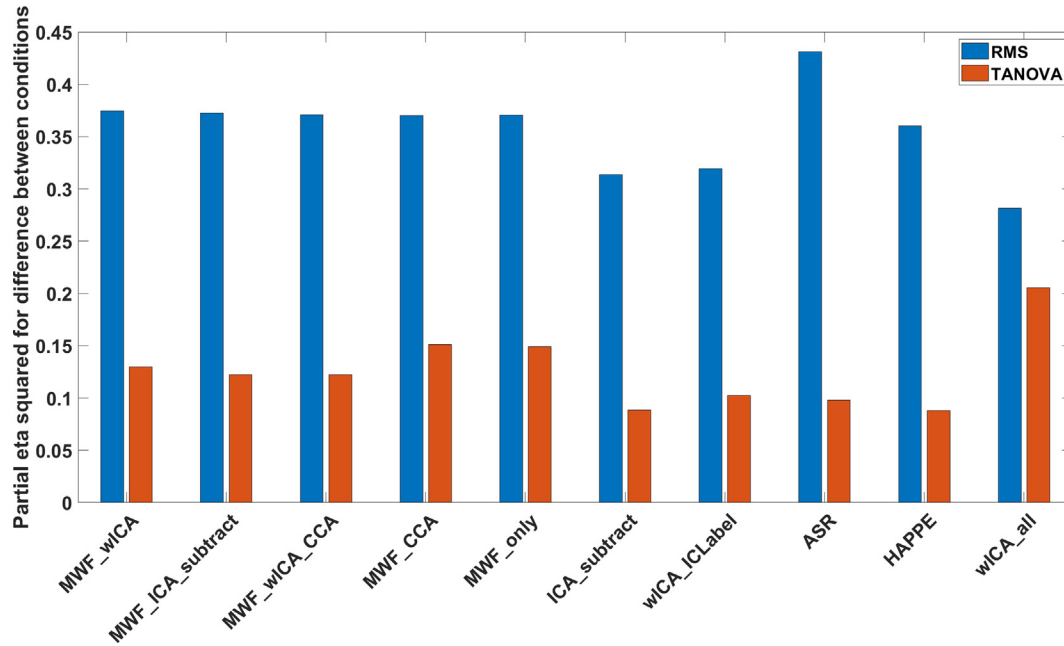
N.W. Bailey, M. Biabani, A.T. Hill et al.

Fig. 8. The variance explained (np^2) by differences in averaged alpha power between eyes open and eyes closed for root mean squared (RMS) and topographical analysis of variance (TANOVA) tests for each of the cleaning pipelines. Please refer to Tables 1 and 2 for an explanation of the labels of the EEG cleaning pipelines presented in this figure.

Together, these findings demonstrate that different cleaning pipelines can alter the EEG signal in differing ways. While the pattern of change during different resting-state and task conditions is broadly similar across pipelines, the choice of pipeline can alter the overall effect size between conditions. Additionally, while most pipelines showed a very similar distribution of alpha activity, HAPPE and wICA_all showed a different distribution of activity to all other pipelines, suggesting (alongside the alpha RMS, SER, and ARR values) that these two pipelines may be overcleaning the data, removing alpha power from specific electrodes such that the distribution of activity after cleaning is significantly altered (see Figs. 11 and 12, and Supplementary Materials Figures S31-32, pages 51–52).

4. Discussion

This study reports the results of comparisons between multiple variations of our newly developed RELAX EEG cleaning pipeline and six other commonly used pipelines. These comparisons were made across a wide range of cleaning and signal preservation metrics. Considering the results of the comparisons across the cleaning pipelines, across multiple datasets with different recording parameters, and taking account of the overall performance on both a range of cleaning quality and signal preservation metrics, we conclude that the RELAX MWF_wICA pipeline provided the best performance at cleaning the data (without "overcleaning" the data), while also performing well at detecting experimental effects. Additionally, either the RELAX MWF_wICA or wICA_ICLabel methods were amongst the highest performing pipelines in metrics assessing the ability to detect experimental effects after cleaning with the pipeline. MWF_wICA also provided high values for both signal preservation (SER) and overall artifact reduction (ARR), with simultaneously higher values for both SER and ARR than ASR or ICA_subtract across all datasets, indicating MWF_wICA was better at both removing artifacts and preserving signal. While MWF_only or wICA_ICLabel approaches maintained more of the signal as indicated by higher SER values, these methods were also less effective at reducing artifacts, as indicated by lower ARR values and poorer

performance in the blink and muscle metrics. Blink and muscle metrics also showed that MWF_wICA was better at cleaning these artifacts than MWF_only, wICA_ICLabel, ICA_subtract, and ASR. Additionally, the RELAX MWF_wICA methods resulted in the least amount of data having to be removed by cleaning and outlying epoch exclusion, except for HAPPE and wICA_all (which our results indicate excessively reduce the neural signal as well as the artifacts). This means that the RELAX MWF_wICA method makes more data available for inclusion in analyses than most pipelines, reducing the risk that participant data needs to be excluded for having too few epochs for inclusion, or that results might be biased by exclusion of epochs containing neural activity due to inferior cleaning. Our companion article additionally indicated that MWF_wICA produced data that showed amongst the highest scores for reliability of ERPs across trials and participants, suggesting cleaning with MWF_wICA provided amongst the most consistency in detecting experimental effects (Bailey et al., 2022). The inclusion of more high-quality epochs for analysis and more reliable data has been suggested by simulations to improve study power even more than increasing sample size (Clayson et al., 2021b, Kolossa and Kopp, 2018, Luck et al., 2021). For these reasons, we recommend the use of the RELAX pipeline with MWF_wICA as a default effective EEG cleaning pipeline when no rationale exists to prefer another variation of RELAX. However, in specific circumstances, the RELAX pipeline with the wICA_ICLabel setting might be preferred (discussed in more detail below). Our results also recommend MWF_wICA or wICA_ICLabel should be preferred over the approach that is most used in the literature currently - ICA_subtract.

Our results indicate the value of combining multiple artifact reduction approaches in a single EEG cleaning pipeline, which is likely to address artifacts that a single method in isolation might not clean completely. For example, muscle activity often appeared to be inadequately addressed by MWF alone, and blinks were not as effectively addressed by wICA_ICLabel alone, while the combination of both methods did effectively clean both artifacts (note that it is easy to perform only MWF or only wICA_ICLabel cleaning with RELAX, so users can test this on their own data). This is in
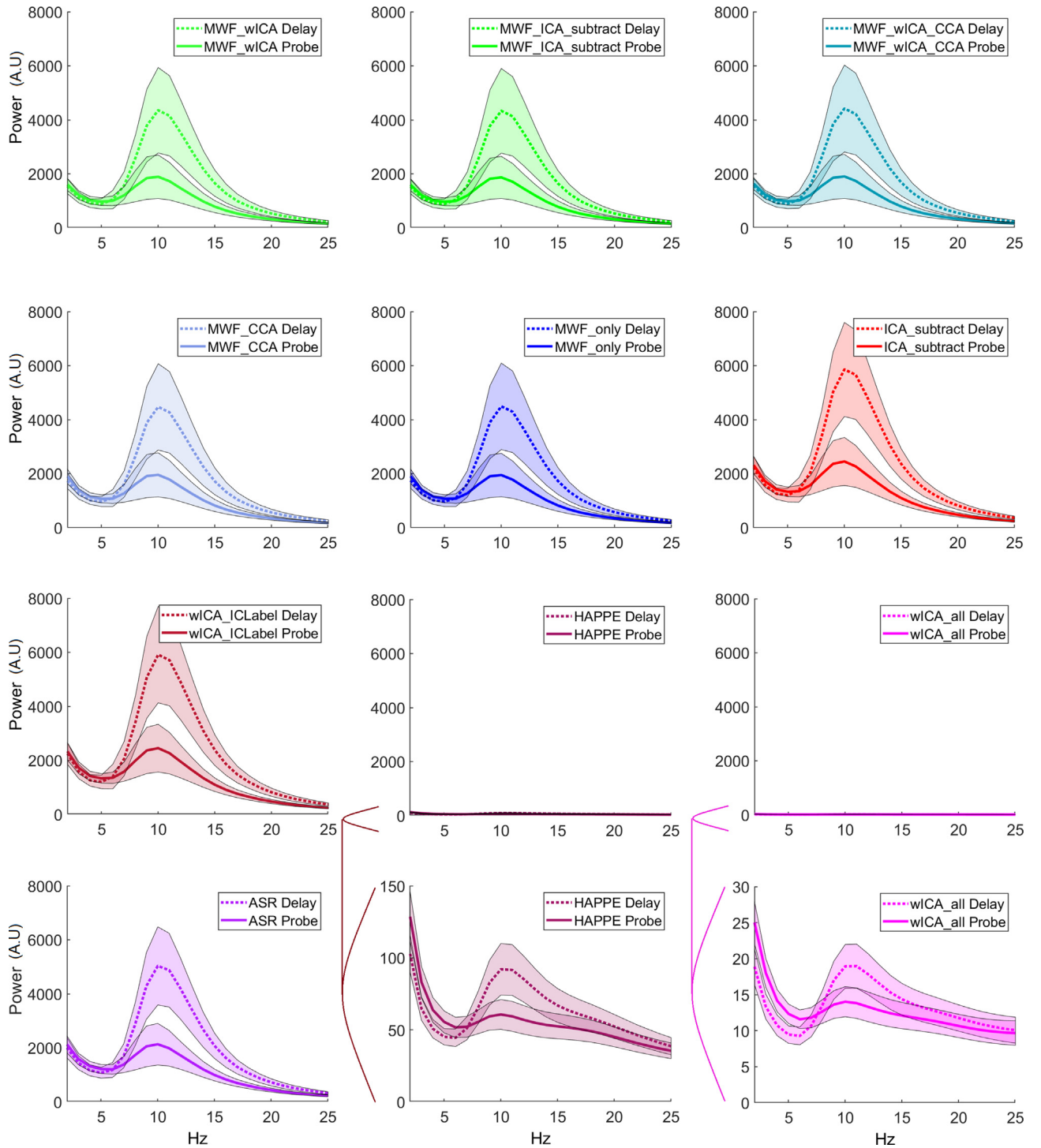
N.W. Bailey, M. Biabani, A.T. Hill et al.

Fig. 9. Frequency power plots depicting Morlet wavelet transformed power averaged across PO7 and PO8 at each frequency from the working memory delay and probe periods for each cleaning pipeline (shaded errors reflect 95% confidence intervals). Note the different scale required to see the power spectrum for HAPPE and wICA_all. Note also that because Morlet wavelet transform power is dependent on the settings used, power values are defined in "arbitrary units" (a.u.). Please refer to Tables 1 and 2 for an explanation of the labels of the EEG cleaning pipelines presented in this figure.

alignment with some other research, where combination cleaning approaches that performed CCA, ICA and independent vector analysis have also been demonstrated to be more effective than single approaches (Barban et al., 2021). We suspect a benefit of the combined MWF and wICA approach is that the combination provided a useful backup (in case one step inadequately addressed an artifact, the other step was still likely to reduce that artifact) while not over-cleaning the data (if the MWF cleaning completely addressed

the blink artifact, ICLabel would not detect any blink artifact and thus the wICA step would not further attempt to reduce blink artifacts leading to over cleaning). However, while adding CCA to the MWF pipeline did improve the cleaning of muscle activity compared to MWF_only, it also resulted in inferior blink correction compared to MWF_wICA, suggesting that wICA applied to artifacts detected by ICLabel is an optimal second step as it addresses all artifacts rather than just muscle artifacts. Additionally, using CCA
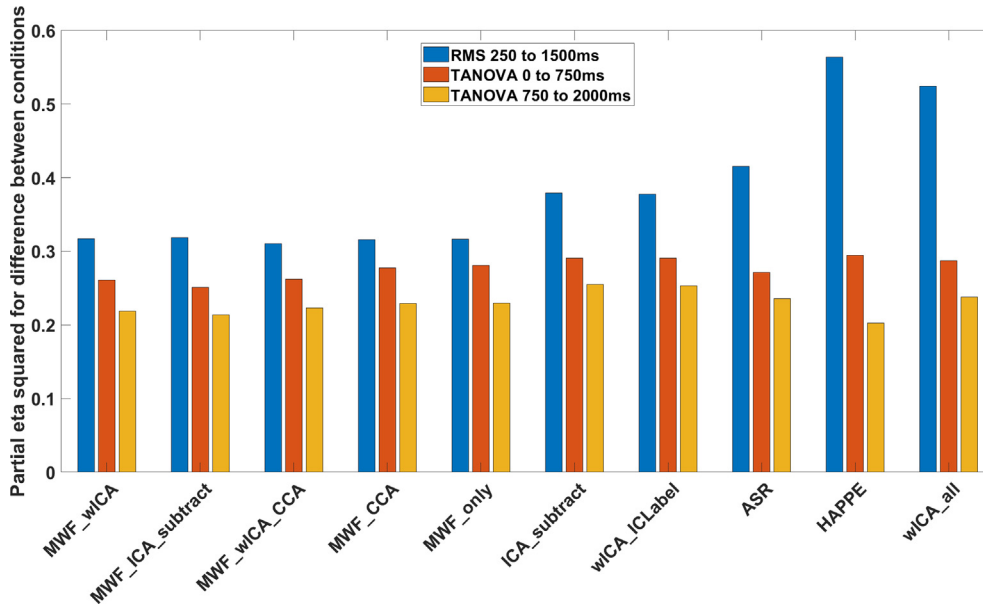
Fig. 10. The variance explained (np²) by the difference in root mean squared (RMS – a measure of overall neural response strength) and topographical analysis of variance (TANOVA – a measure of differences in the distribution of neural activity) tests comparing alpha activity between the delay and probe periods of the Sternberg task after cleaning by each pipeline. Please refer to Tables 1 and 2 for an explanation of the labels of the EEG cleaning pipelines presented in this figure.

in addition to MWF_wICA (MWF_wICA_CCA) did not enhance performance compared to MWF_wICA. As such, while the application of specific multiple cleaning methods does seem to improve cleaning performance, it is not the case that adding more cleaning methods is always better. Additionally, our results suggested that both when applied in isolation and after MWF cleaning, wICA_ICLabel was better at preserving signal (higher SER values) than ICA_subtract, while still providing similar artifact reduction (ARR values). Since the SER reflects the ability of a pipeline to preserve signal in periods that are not affected by artifacts, we suspect this highlights the manner in which subtracting an ICA component means removing not only artifact, but neural activity that is mixed into the component due to imperfect ICA decomposition. In contrast, because of the wavelet thresholding used within wICA, only the largest amplitude signal is removed by the wICA_ICLabel method (and this largest amplitude signal from a component identified as an artifact is likely to reflect the artifact, whereas the smaller amplitude signal that is left in the data is more likely to reflect neural activity), while the clean data periods (which do not contain this larger amplitude signal) are left unaffected (allowing wICA_I-CLabel to produce higher SER values). Similarly, while subtraction of ICA after MWF reduced the metrics assessing remaining muscle activity to approximately zero, the method never led to higher explained variance from the experimental tests and resulted in lower variance explained in most metrics. As such, while subtracting artifactual components with ICA is probably the most common cleaning approach, we recommend the use of wICA to reduce artifactual components instead. Given these two points, our recommended RELAX pipeline applies MWF cleaning then wICA cleaning (but not CCA or ICA subtraction).

While the preceding paragraphs suggest that the cleaning metrics we tested strongly recommend the use of RELAX with the MWF_wICA setting, our perspective is that the most important metric for cleaning efficacy is whether a cleaning approach enhances a researcher's ability to detect experimental outcomes (a view shared by other researchers) (Clayson et al., 2021a). Here, the results do not indicate a single best cleaning approach for all applications. This is consistent with previous research, which has also been unable to recommend a single best pipeline for all situations (Barban et al., 2021). Encouragingly, except for the wICA_all

and HAPPE pipelines, all the cleaning approaches we tested led to broadly similar amounts of variance explained by the experimental manipulation and similar patterns for outcome measures. This suggests that as long as the confounding influences of artifacts are sufficiently addressed and signal is sufficiently preserved, EEG cleaning choices may not severely confound results across different studies, at least providing effects in the same direction and of similar magnitudes. However, we deliberately chose tasks that provided well established reasonably sized differences between the experimental conditions. Studies examining differences between clinical groups or treatment conditions might show smaller effect sizes than between condition comparisons, in which case optimal pipeline selection might be more likely to influence results. Indeed, some research has suggested that subtle effects do not always replicate between different data cleaning approaches, and that effect sizes vary across EEG cleaning pipelines (Clayson et al., 2021a, Robbins et al., 2020, Rogasch et al., 2020). As such, we recommend the use of RELAX MWF_wICA as a default due to its superior artifact cleaning and signal preservation attributes. However, for specific uses, the wICA_ICLabel setting might at times be preferred.

In particular, the amount of variance explained by the experimental manipulations indicated that the MWF_wICA pipeline did not significantly differ from the most effective pipelines when examining resting-state alpha power. However, the wICA_ICLabel setting of RELAX performed more highly in alpha power measures from the Sternberg task, and as such might be preferred for studies focused on similar oscillatory outcome measures. Having said that, because wICA_ICLabel did not clean blinks or muscle as effectively, we recommend wICA_ICLabel only be used where enough trials are collected (or enough participants) that lower reliability could be acceptable, as research has suggested data quality is more important for power than data quantity (Kolossa and Kopp, 2018). As such, wICA_ICLabel might be preferred when using robust statistics and methods that include all single trials in the statistical inferences, where reduced cleaning can avoid the potential negative effects of arbitrary cleaning thresholds, and robust statistics can account for the noisier data (Alday and van Paridon, 2021). However, note that if an inadequately cleaned artifact is time-

N.W. Bailey, M. Biabani, A.T. Hill et al.

Fig. 11. Alpha power distribution during the early period (0 to 750 ms) after the stimuli of the working memory delay (retention) and working probe periods from each of the cleaning pipelines. All plots are on the same Morlet wavelet transform power scale so they can all be compared to all other pipelines [note that because Morlet wavelet transform power is dependent on the settings used, power values are defined in "arbitrary units" (a.u.)]. Note that wICA_all and HAPPE reduced the power in all frequencies by approximately 2 orders of magnitude. Please refer to Tables 1 and 2 for an explanation of the labels of the EEG cleaning pipelines presented in this figure.

locked to a period of interest, even robust statistics might not address its potential to confound results.

wICA_ICLabel might additionally be recommended for studies examining connectivity between electrodes, as the effect of MWF on connectivity measures is currently unknown. For researchers interested in connectivity measures, it may be useful to note that cleaning EEG data using wavelet enhanced ICA (wICA) to reduce artifacts does not reduce the rank of the data, so might allow for higher resolution of nodes when using connectivity analysis in source space (in contrast to subtracting ICA components, which does reduce the rank) (Castellanos and Makarov, 2006). Additionally, for research examining gamma oscillations, it might be worth considering that MWF_ICA_subtract removed essentially all muscle activity, which is a potential confound in studies examining gamma oscillations due to the overlap in frequencies between muscle activity and gamma neural oscillations. However, note that the 'ground-truth' of gamma oscillations was not determined, so we cannot eliminate the possibility that MWF_ICA_subtract removed neural oscillations as well as muscle.

In contrast to the preceding recommendations, while ASR provided high values for the variance explained between our experimental conditions, there were no cases where it provided higher performance than both wICA_ICLabel and MWF_wICA, and ASR performed worse than MWF_wICA in all artifact reduction metrics.

Our results also suggested high variability in some measures after ASR cleaning, and our experience using ASR has suggested it produced inconsistent cleaning outcomes even when applied to the same file (including a high number of periods removed as reflecting extreme artifacts when using the default settings). Unfortunately, we were not able to discern whether this variability reflects 'ground-truth' differences (in which case the variability is valuable), or variability produced by the ASR cleaning (in which case the variability is artifactual), and our objective measures do not address the cause of the variability.

Our results also indicated that HAPPE and wICA_all were associated with very low SER values, and very low amplitude EEG activity after cleaning (including a reduction by > 2 orders of magnitude for alpha power), indicating considerable reduction of the neural activity signal. After cleaning with these pipelines, the majority of epochs also showed log-power log-frequency slopes that were indicative of residual muscle activity. It may be that this was not because these pipelines inadequately cleaned muscle activity, but rather they over-cleaned the neural activity signal in the lower frequency ranges, so the cleaned slopes were flattened and no longer displayed the 1/f power-frequency distributions of typical neural activity (Donoghue et al., 2020). However, our power frequency plots for wICA_all and HAPPE pipelines did show more power in the beta frequency range relative to lower frequencies compared
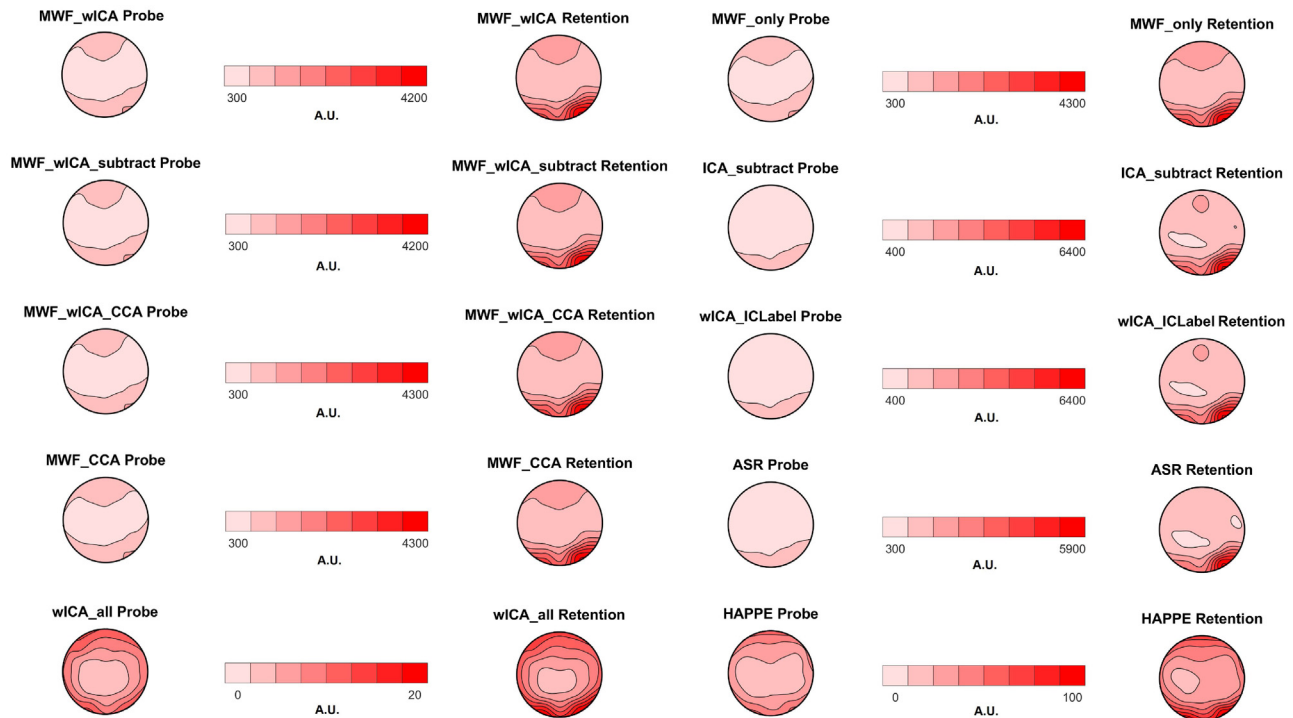
Fig. 12. Alpha power distribution during the early period (0 to 750 ms) after the stimuli of the working memory delay (retention) and working probe periods from each of the cleaning pipelines. All pipelines are on their own individual Morlet wavelet transform power scale so the pattern of alpha activity distributions can be viewed within each pipeline (but comparisons of amplitudes between pipelines are not meaningful in this figure). Note that because Morlet wavelet transform power is dependent on the settings used, power values are defined in "arbitrary units" (a.u.). As expected, the retention period showed more occipital / parietal alpha maximums, while the probe period showed more widespread alpha. Note the similarity in pattern across most pipelines, including ICA only and MWF only methods, implying different cleaning approaches still lead to similar patterns. ASR, HAPPE, and wICA_all were the most different to the other cleaning pipelines, with ASR removing most of the frontal alpha, and HAPPE and wICA showing less qualitative differentiation between the probe and retention period alpha, despite these pipelines showing the highest variance explained for the comparison between the two conditions. We suspect this may be the result of removal of significant amount of the variance by these pipelines, leading to highly precise estimates of the retention and probe alpha activity, which allow strong inferences about the differences in these periods (which no longer overlap after the removal of so much variance). However, we suspect this comes at the cost of distorting the distribution/characterization of the neural activity. Please refer to Tables 1 and 2 for an explanation of the labels of the EEG cleaning pipelines presented in this figure.

to other pipelines (Figs. 7 and 9), so it may be that these methods simply do not adequately clean muscle activity. The combination of low SER values, reduced power frequency amplitudes, and altered activity distributions compared to the other pipelines suggests that the wICA_all and HAPPE pipelines overclean the data, removing elements of the neural signal. It is worth noting that HAPPE was designed to address noisy data with large artifacts, so our result perhaps highlights that it is not appropriate for typical cognitive data without excessive noise (Gabard-Durnam et al., 2018). As such, while the HAPPE and wICA_all methods did seem to produce higher explained variance for several experimental outcomes, we do not recommend their use, except perhaps in the case of extremely noisy data, or where the outcome of interest is not concerned with characterizing the data, but simply differentiating two conditions (for example, brain-computer interface applications might usefully apply the wICA_all approach in real time using very rapid ICA computation approaches).

Since it seems there is not a single best cleaning pipeline that can be recommended, we have made RELAX modular so that different approaches can easily be implemented. Some default settings can be recommended. For example, the results reported in our companion paper indicated that the ICA methods infomax or fastICA with the symmetric setting were the best performers within RELAX (Bailey et al., 2022). If users can install cudaICA, this setting is likely to be optimal, providing both speed and the best performance (Raimondo et al., 2012). Traditional extended infomax can also be used, but is slower (Lee et al., 1999). Otherwise, we have

left the default setting as fastICA with the symmetric setting (Hyvarinen, 1999), which our companion paper has indicate is (just slightly) better than the deflation setting (Bailey et al., 2022). We have set this fastICA to repeat the ICA up to three times in the case of a non-convergence issue which can adversely affect ICA decompositions, then to switch to the 'defl' setting if non-convergence still occurs. Other defaults have been set after an extensive process of informal testing to determine optimal cleaning and values for the variance explained by different experimental manipulations. If deviating from these defaults, then specific RELAX settings or parameter variations used in future research should be justified, primarily with reference to a previous demonstration of maximal explained variance in a similar experimental design. Ideally, the parameters selected should also be pre-registered to prevent "fishing" for positive results. However, while we recommend the use of methods that ensure "fishing" does not inflate the false positive rate, it may also be useful to other researchers to report analyses using multiple parameter selections. This would help guide future research towards the optimal methods for particular use cases. It may have the additional benefit of demonstrating robustness of results to variations in cleaning methods, eliminating the possibility that a specific result is produced by or dependent on a specific cleaning method. If researchers implement this approach, the a priori selected, primary, and ideally pre-registered cleaning method should be specified so readers can be aware of which analysis method was pre-planned, and understand the risk that positive results were produced by multiple re-analyses. RELAX provides

an output of the cleaning efficacy metrics used in the current study, which should be reported so reviewers and readers can determine the risk that remaining artifacts could confound conclusions (perhaps in a Supplementary Materials section for the sake of brevity). We have provided further discussion points in our Supplementary Materials, section 7, pages 75–76).

4.1. Limitations

Despite our extensive testing, there are many limitations to the conclusions we can draw from the current study. Firstly, while our results are relatively comprehensive, we have only scratched the surface of the full potential parameter space of cleaning options. In particular, we note that we did not make comparisons with signal space projection approaches to cleaning EEG data (Tesche et al., 1995, Uusitalo and Ilmoniemi, 1997), such as the source-estimate-utilizing noise-discarding algorithm (SOUND) (Mutanen et al., 2018). We have also only tested the cleaning pipelines on a small number of datasets (although certainly with large sample sizes), and only tested the cleaning pipelines on a minority of experimental tasks and experimental outcome measures (although more than most cleaning studies - see also our companion article for examination of ERP related data, (Bailey et al., 2022)). In particular, we did not test the effect of data cleaning approaches on connectivity analyses or metrics that assess the 1/f aperiodic non-oscillatory component of the data.

Secondly, we have not tested RELAX on extremely noisy data contaminated with large movement artifacts, nor scans of children/infants, nor the elderly who often produce EEG data with more artifacts, and which may contain different characteristics to the datasets we tested, potentially requiring alternative artifact/signal discrimination approaches. We note that ICLabel was built sampling from adult data (Pion-Tonachini et al., 2019), and the log-power log-frequency muscle slope thresholds were derived from adult data (Fitzgibbon et al., 2016). Similarly, ICLabel was built from data containing 32 or more electrodes, so we suspect data with less than 32 electrodes might not provide ICLabel with sufficient data for optimal artifact detection, leading to reduced cleaning performance (Pion-Tonachini et al., 2019). All of our test data also included between 44 to 62 electrodes, so the efficacy of RELAX at cleaning high-density EEG montages with > 100 electrodes is not certain. While the principles used to exclude extremely bad epochs and channels should still apply, it may be that with denser montages a higher proportion of data periods contain at least a single electrode that exceeds the threshold suggesting the data period should be rejected. The methods used to identify artifacts for MWF cleaning should not differ in denser EEG montages, but the MWF cleaning will require more memory and computational power. Additionally, ICLabel was trained on data up to 128 electrodes, so should be able to be implemented on denser EEG montages (Pion-Tonachini et al., 2019). It is also worth noting that short EEG data periods may not have enough samples for effective ICA computation, nor for identification of artifacts for MWF cleaning. At least [Number of channels ^ 2 × (~30)] data points are recommended (~2.5 minutes for 64 channel data recorded at 1000 Hz) (Miyakoshi, 2018). Additionally, our test datasets all used an electrode near to CPz as the online reference, and re-referencing was not performed until after the MWF cleaning. Our preliminary tests suggested that the MWF cleaning step was slightly more effective when CPz remained the reference electrode than when data were re-referenced to the average electrode. Preliminary testing indicated RELAX did work with a linked ear referencing montage, but we did not formally test cleaning using other referencing montages. As such, if a different recording montage has been used, we would suggest re-referencing to CPz or a nearby electrode prior to submitting data for cleaning with RELAX (remembering that data cleaned with RELAX is outputted with the robust common average reference implemented, and that data can be then re-referenced to a user's chosen referencing montage if desired).

Third, while automation of EEG data pre-processing prevents the risk that subjectivity would affect the data compared to manual pre-processing, there are also potential pitfalls. One particularly risky pitfall is that some files may be corrupted during the recording process and contain no neural related data whatsoever, an issue that automated cleaning methods may not identify. This issue occurred for one file we submitted to our pre-processing pipelines. As such, we recommend that EEG recording technicians exclude data files that appear to contain no usable signal before submitting files to pre-processing. As a double-check, we have also included in the pipeline a visual test of outlying median voltages from each electrode in the cleaned data, which indicates if a particular file contains atypical data which should be manually checked (a method that effectively identified our problem file).

Additionally, there is no way to be sure of the 'ground-truth' of our tests of the amount of variance explained by different experimental manipulations. As such, it may be possible that the higher levels of explained variance are due to artifacts rather than actual differences in neural activity. However, this seems an unlikely explanation for our results for a number of reasons. First, the artifacts would have to be both time-locked to the stimuli in the case of the Sternberg task (and the Go Nogo data reported in our companion manuscript) and appear more strongly or more frequently in one condition than the other. The only artifact that is likely to fit this requirement is eye movement. However, all pipelines produce blink amplitude ratios ranging from 1.015 to 1.077, values that are likely too small to produce much influence on differences in variance explained between pipelines. Second, we think it is unlikely that any artifact would increase the explained variance by the experimental manipulation across multiple different tasks and types of brain activity (especially considering the different topographical patterns and underpinning mechanisms associated with each measure). We suggest this is a particularly unlikely explanation, since the measures used were selected because they have been robustly characterised by previous research, which demonstrated that the experimental manipulations lead to the patterns of brain activity that we detected in the current study. Finally, if the explained variance was due to the influence of artifacts, we would instead see an artifact pattern in the TANOVA for one of the conditions from one of the cleaning pipelines (the most obvious of which would have been a blink topography). No such artifact topography was present in any of our pipelines or comparisons for the tests of experimental effects. While it may be that previous research that has demonstrated the expected effect of the different experimental manipulations might have also been driven by artifacts in the data, these artifacts would have to be consistent across the range of cleaning pipelines that have replicated the effect for the analyses we included (including the range of pipelines tested in our study, which almost all showed the expected effects). Furthermore, in some cases, previous research has used electrocorticography or fMRI in conjunction with EEG to demonstrate that EEG measures relate to brain activity detected with these other methods, providing confidence that artifacts are not producing the result (Ahmad et al., 2016, Baumeister et al., 2014, Iannaccone et al., 2015, Meltzer et al., 2007, Smith et al., 2013, Zhang et al., 2018). The putative brain activity detected for many of the metrics we included has also been shown to predict behavioural performance, suggesting functional relationships that are unlikely to be present for artifacts (Bashivan et al., 2014, Clark et al., 2004, Karamacoska et al., 2018). As such, we suggest that our results indicate that RELAX cleans artifacts in such a way that the signal (brain activity) to noise (muscle, eye movement etc.) ratio is maximised, and the higher explained variance values reflect better detection of the effects of the experimental manipu-

lation on the brain activity rather than artifacts remaining after cleaning, providing a strong recommendation for the use of RELAX when cleaning EEG data.

In order to address this 'ground-truth' issue, one potential recommendation could be to use simulated data that includes artifacts. One method for adding artifacts to a ground truth-signal that has been implemented in the past is to obtain an artifact signal that has been removed by cleaning real data, then adding this artifact signal to simulated neural data (Hoffmann and Falkenstein, 2008). Unfortunately, this can create a bias towards the cleaning method used to remove those artifacts from the real data (Barban et al., 2021, Hoffmann and Falkenstein, 2008, Urigüen and Garcia-Zapirain, 2015). To avoid this bias, previous research has recommended testing the cleaning efficacy of different methods on multiple versions of simulated data, constructed from the use of multiple methods for extracting the artifact templates (Hoffmann and Falkenstein, 2008). However, this approach might simply provide multiple biased assessments of cleaning efficacy. As such, another approach is to use real artifacts as templates to create artifact models that are distributed in a spatio-temporally realistic manner, then add these simulated artifacts (that are independent from the cleaning methods being tested) to the ground-truth data (Barban et al., 2021). Unfortunately, regardless of the method used, these simulated data construction methods still do not preserve certain aspects of real data - for example, the synchronisation between neural activity and stimulus presentation, synchronisation between electrodes (Urigüen and Garcia-Zapirain, 2015), or, perhaps most critically for the testing of effective EEG cleaning methods, synchronisation between the neural activity that elicits blink or muscle artifacts and the timing of those artifacts. As such, while real data (as tested in the current study) provides the most important final test of any cleaning method, simulations may be an effective method to obtain otherwise biologically plausible artifacts with a ground-truth neural signal enabling researchers to assess the match between cleaning outcomes and the ground-truth data (Urigüen and Garcia-Zapirain, 2015).

4.2. Potential improvements and future research

We tested many potential versions of RELAX, and informally explored most parameter settings for optimal performance prior to formal testing. However, there are potential improvements that may be possible which are worth exploring in future EEG cleaning pipeline development. These are discussed in full in the Supplementary Materials (section 7, page 76), but include: 1) using an adaptive threshold for the wICA cleaning (which we briefly tested using an array of the level dependent settings within the MATLAB function 'wdenoise', but found this approach to be less effective at reducing blink artifacts); 2) using a less stringent amplitude threshold for outlying epoch rejection after cleaning, or alternatively, no epoch rejection and robust statistics for comparisons (Alday and van Paridon, 2021); 3) taking temporal information into account in the ICA computation (with Independent Vector Analysis for example; (Barban et al., 2021), although we note that this is slow and potentially complicated to implement); 4) adapting ICLabel to identify artifactual components based on empirically established objective methods, rather than its current design which used expert consensus [for example, applying the objective approach to identifying muscle components based on comparison of paralysed vs non-paralysed scalp recordings provided by Fitzgibbon et al. (2016)]; and 5) RELAX was designed for effective cleaning without the need for electrocardiogram or electrooculogram electrodes to maximise its applicability and consistency (in our experience, these electrodes are more commonly noisy than scalp electrodes). However, the inclusion of electrocardiogram or electrooculogram electrodes may allow more accurate identifica-

tion of the eye movement or cardiac artifacts and is worth exploring in future research.

The current study also required a very large amount of computation and researcher hours, involving massive repeated iterative testing of different cleaning parameters by multiple researchers across multiple datasets and multiple outcome metrics. As such, the comprehensiveness of a project like this is difficult to achieve. Despite this difficulty, we would recommend examining an even higher number of metrics and varying an even broader parameter space in future pipeline development. To achieve this, we suggest to the field that an online resource containing multiple large EEG datasets could be collated for the purposes of testing EEG cleaning pipelines (note that small scale examples that have been set up to enable testing of cleaning approaches already exist (Barban et al., 2021, Kappenman et al., 2021, Zhang et al., 2020), and large-scale open access datasets are also available (van Dijk et al., 2022). The cleaning efficacy of newly proposed pipelines could be compared via these datasets across multiple metrics, and after cleaning, multiple neural activity measures of explained variance could be assessed (for example, ERPs, oscillations, 1/f metrics, and connectivity measures). A 'leader board' of the current optimal methods could be established, so that new cleaning approaches would not have to replicate pipelines that have already been tested to make comparisons. An automated algorithm could iteratively test parameter details that may affect cleaning quality (such as the wICA threshold or outlier identification thresholds). It is also possible for different settings of different parameters to interact with each other to influence cleaning performance. For example, lower extreme outlier exclusion thresholds may improve wICA performance more for higher wICA thresholds than lower thresholds. As such, the potential parameter space might be practically infinite. This being the case, automated EEG cleaning pipeline development is much better performed by an automated system, perhaps with many pipelines being tested concurrently to optimise pipeline development as quickly as possible. If this were achieved, the most optimal current pipeline could be automatically uploaded to GitHub each time the testing system discovered an improvement. Different outcome measures or experimental designs could even have their own version of the most optimal pipeline (Clayson et al., 2021a). As such, the system could reflect a "living" gold-standard in EEG pre-processing. We unfortunately do not have the resources to implement this approach, and we are not able to share our data due to ethical approval criteria. We mention it here in the hope that the field may be inspired to achieve this, perhaps via large scale collaboration (we note that #EEGManyLabs or #EEGManyPipelines could be good platforms for this discussion) (Algermissen et al., 2021, Pavlov et al., 2021).

5. Conclusion

To conclude, we recommend future researchers use the RELAX approach with the default MWF_wICA setting for cleaning EEG data when no clear rationale exists to use another setting, as it provides the following benefits:

1) RELAX MWF_wICA provided the best or equivalent to the best cleaning of all artifact types.
2) RELAX is fully automatic and quick to apply, saving considerable time, and eliminating subjectivity throughout the cleaning process.
3) The RELAX pipeline contains embedded cleaning quality metrics, which can also be easily reported in publications. This will enable researchers to evaluate the likelihood that the specific cleaning parameters or remaining artifacts might be influencing conclusions drawn from the data.

4) The RELAX MWF_wICA approach maximises the number of epochs included for analysis, thus reducing the possibility of biasing the data based on epoch exclusion and increasing the statistical power of the study.

5) RELAX MWF_wICA provided high levels of variance explained by the experimental manipulation.

We have provided the source code and made the pipeline straightforward to use and publicly accessible, with the only technical skills required being the installation of the MATLAB specific and external dependencies. RELAX also does not require eye electrodes (unlike regression approaches), so is applicable to a wider range of datasets and robust against participant discomfort with eye electrodes (Gómez-Herrero et al., 2006), as well as the somewhat common (in our experience) "bad eye electrode" issue. RELAX is available for download from GitHub (https://github.com/NeilwBailey/RELAX/releases). Instructions for installing and using the pipeline can be found in the GitHub wiki: https://github.com/NeilwBailey/RELAX/wiki. The pipeline is designed to run within the commonly used EEGLAB software (implemented in MATLAB) (Delorme et al., 2007). Finally, we note the acronym "RELAX" was also chosen as a prompt for researchers - while technological advances have the potential to reduce our workload, they have typically made us simply more productive instead. Given the stressors researchers are exposed to (Bowen et al., 2016), and the reductions in health associated with that stress (Faragher et al., 2005, Nixon et al., 2011), we recommend that at least part of the time saved by using our pipeline could be spent on researcher well-being, rather than simply further increasing productivity.

Recommendations for using RELAX when examining oscillations:

1. Use RELAX_MWF_wICA as the default cleaning pipeline (with fastICA on the symm setting, or cudaICA if it can be installed)

2. Use RELAX_wICA_ICLabel if analysing task related oscillatory power or connectivity, but only if data are relatively clean, if using analysis methods that prevent remaining artifacts confounding conclusions (for example single trial analyses with robust statistics) and if a large number of trials is available for analysis.

3. Consider using RELAX_MWF_ICA_subtract if analysing gamma power (or if removing all muscle activity is required)

   See our companion article for recommendations related to ERP analyses (Bailey et al., 2022).

Declaration of Interest

PBF has received equipment for research from MagVenture A/S, Nexstim, Neuronetics and Brainsway ltd and funding for research from Neuronetics. He is a founder of TMS Clinics Australia and Resonance Therapeutics. NWB, MB, ATH, AM, NCR, BM, and OWM have no competing interests to declare.

Role of Funding Source

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.clinph.2023.01.017.

References

Ahmad RF, Malik AS, Kamel N, Reza F, Abdullah JM. Simultaneous EEG-fMRI for working memory of the human brain. Australas Phys Eng Sci Med 2016;39(2):363–78.

Akhtar MT, Mitsuhashi W, James CJ. Employing spatially constrained ICA and wavelet denoising, for automatic removal of artifacts from multichannel EEG data. Signal Process 2012;92(2):401–16.

Alday PM, van Paridon J. Away from arbitrary thresholds: using robust statistics to improve artifact rejection in ERP. PsyArXiv Preprints 2021. https://doi.org/10.31234/osf.io/wqrb5.

Algermissen J, Yang Y, Busch N. EEGManyPipelines: Mapping the diversity of EEG analysis pipelines and their impact on results. Retrieved from osf.io/xfrbe 2021.

Allen M, Poggiali D, Whitaker K, Marshall TR, Kievit RA. Raincloud plots: a multi-platform tool for robust data visualization. Wellcome Open Res 2019;4:63.

Anders M, Anders B, Kreuzer M, Zinn S, Walter C. Application of referencing techniques in EEG-based Recordings of Contact Heat Evoked Potentials (CHEPS). Front Hum Neurosci 2020;14:527.

Bailey N, Hill A, Biabani M, Murphy O, Rogasch N, McQueen B, et al. Introducing RELAX (the Reduction of Electroencephalographic Artifacts): A fully automated pre-processing pipeline for cleaning EEG data – Part 2: Application to Event-Related Potentials. Clin Neurophysiol 2022. this volume.

Bailey NW, Freedman G, Raj K, Spierings KN, Piccoli LR, Sullivan CM, et al. Mindfulness meditators show enhanced accuracy and different neural activity during working memory. Mindfulness 2020;11:1762–81.

Barban F, Chiappalone M, Bonassi G, Mantini D, Semprini M. Yet another artefact rejection study: an exploration of cleaning methods for biological and neuromodulatory noise. J Neural Eng 2021;18(4). https://doi.org/10.1088/1741-2552/ac01fe.

Bashivan P, Bidelman GM, Yeasin M. Spectrotemporal dynamics of the EEG during working memory encoding and maintenance predicts individual behavioral capacity. Eur J Neurosci 2014;40(12):3774–84.

Baumeister S, Hohmann S, Wolf I, Plichta MM, Rechtsteiner S, Zangl M, et al. Sequential inhibitory control processes assessed through simultaneous EEG-fMRI. Neuroimage 2014;94:349–59.

Bender R, Lange S. Adjusting for multiple testing—when and how? J Clin Epidemiol 2001;54(4):343–9.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statistical Soc: Ser B (Methodol) 1995;57(1):289–300.

Bertrand A. Distributed signal processing for wireless EEG sensor networks. IEEE Trans Neural Syst Rehabil Eng 2015;23(6):923–35.

Bigdely-Shamlo N, Mullen T, Kothe C, Su K-M, Robbins KA. The PREP pipeline: standardized preprocessing for large-scale EEG analysis. Front Neuroinform 2015;9:16.

Borowicz A. Using a multichannel Wiener filter to remove eye-blink artifacts from EEG data. Biomed Signal Process Control 2018;45:246–55.

Bowen P, Rose R, Pilkington A. Perceived stress amongst university academics. Am Int J Contemporary Res 2016;6(1):22–8.

Castellanos NP, Makarov VA. Recovering EEG brain signals: artifact suppression with wavelet enhanced independent component analysis. J Neurosci Methods 2006;158(2):300–12.

Chang C-Y, Hsu S-H, Pion-Tonachini L, Jung T-P. Evaluation of artifact subspace reconstruction for automatic artifact components removal in multi-channel EEG recordings. IEEE Trans Biomed Eng 2019;67(4):1114–21.

Clark CR, Veltmeyer MD, Hamilton RJ, Simms E, Paul R, Hermens D, et al. Spontaneous alpha peak frequency predicts working memory performance across the age span. Int J Psychophysiol 2004;53(1):1–9.

Clayson PE, Baldwin S, Rocha H, Larson MJ. The Data-Processing Multiverse of Event-Related Potentials (ERPs): a roadmap for the optimization and standardization of ERP processing and reduction pipelines. Neuroimage 2021a;245 118712.

Clayson PE, Carbine KA, Baldwin SA, Olsen JA, Larson MJ. Using generalizability theory and the ERP Reliability Analysis (ERA) Toolbox for assessing test-retest reliability of ERP scores Part 1: algorithms, framework, and implementation. Int J Psychophysiol 2021b;166:174–87.

Delorme A, Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods 2004;134(1):9–21.

Delorme A, Sejnowski T, Makeig S. Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. Neuroimage 2007;34(4):1443–9.

Dimigen O. Optimizing the ICA-based removal of ocular EEG artifacts from free viewing experiments. Neuroimage 2020;207 116117.

Donoghue T, Haller M, Peterson EJ, Varma P, Sebastian P, Gao R, et al. Parameterizing neural power spectra into periodic and aperiodic components. Nat Neurosci 2020;23(12):1655–65.

Faragher EB, Cass M, Cooper CL. The relationship between job satisfaction and health: a meta-analysis. Occup Environ Med 2005;62(2):105–12. https://doi.org/10.1136/oem.2002.006734.

Fitzgibbon S, DeLosAngeles D, Lewis T, Powers D, Grummett T, Whitham E, et al. Automatic determination of EMG-contaminated components and validation of independent component analysis using EEG during pharmacologic paralysis. Clin Neurophysiol 2016;127(3):1781–93.

N.W. Bailey, M. Biabani, A.T. Hill et al.

Gabard-Durnam LJ, Mendez Leal AS, Wilkinson CL, Levin AR. The Harvard Automated Processing Pipeline for Electroencephalography (HAPPE): standardized processing software for developmental and high-artifact data. Front Neurosci 2018;12:97.

Gómez-Herrero G, De Clercq W, Anwar H, Kara O, Egiazarian K, Van Huffel S, et al. Automatic removal of ocular artifacts in the EEG without an EOG reference channel. In: Proceedings of the 7th Nordic Signal Processing Symposium-NORSIG 2006: IEEE; 2006. p. 130-3.

Habermann M, Weusmann D, Stein M, Koenig T. A student's guide to randomization statistics for multichannel event-related potentials using ragu. Front Neurosci 2018;12:355.

Hoffmann S, Falkenstein M. The correction of eye blink artefacts in the EEG: a comparison of two prominent methods. PLoS One 2008;3(8):e3004.

Hyvarinen A. Fast ICA for noisy data using Gaussian moments. In: 1999 IEEE International Symposium on Circuits and Systems (ISCAS), Orlando, FL, USA, 1999, pp. 57-61 vol. 5.

Iannaccone R, Hauser TU, Staempfli P, Walitza S, Brandeis D, Brem S. Conflict monitoring and error processing: new insights from simultaneous EEG–fMRI. Neuroimage 2015;105:395–407.

Inuso G, La Foresta F, Mammone N, Morabito FC. Wavelet-ICA methodology for efficient artifact removal from Electroencephalographic recordings. In: 2007 International Joint Conference on Neural Networks: IEEE; 2007. p. 1524-9.

Islam MK, Rastegarnia A, Yang Z. Methods for artifact detection and removal from scalp EEG: a review. Neurophysiologie Clinique/Clinical Neurophysiology 2016;46(4–5):287–305.

Issa MF, Juhasz Z. Improved EOG artifact removal using wavelet enhanced independent component analysis. Brain Sci. 2019;9(12):355.

Janani AS, Grummett TS, Lewis TW, Fitzgibbon SP, Whitham EM, DelosAngeles D, et al. Improved artefact removal from EEG using Canonical Correlation Analysis and spectral slope. J Neurosci Methods 2018;298:1–15.

Kappenman ES, Farrens JL, Zhang W, Stewart AX, Luck SJ. ERP CORE: an open resource for human event-related potential research. Neuroimage 2021;225 117465.

Karamacoska D, Barry RJ, Steiner GZ. Electrophysiological underpinnings of response variability in the Go/NoGo task. Int J Psychophysiol 2018;134:159–67.

Kleifges K, Bigdely-Shamlo N, Kerick SE, Robbins KA. BLINKER: Automated extraction of ocular indices from EEG enabling large-scale analysis. Front Neurosci 2017;11:12.

Koenig T, Kottlow M, Stein M, Melie-García L. Ragu: a free tool for the analysis of EEG and MEG event-related scalp field data using global randomization statistics. Comput Intell Neurosci 2011;2011 938925.

Kolossa A, Kopp B. Data quality over data quantity in computational cognitive neuroscience. Neuroimage 2018;172:775–85.

Kumaravel VP, Farella E, Parise E, Buiatti M. Near: An artifact removal pipeline for human newborn EEG data. Dev Cogn Neurosci 2022: 101068.

Lee T-W, Girolami M, Sejnowski TJ. Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. Neural Comput 1999;11(2):417–41.

Luck SJ, Stewart AX, Simmons AM, Rhemtulla M. Standardized measurement error: a universal metric of data quality for averaged event-related potentials. Psychophysiology 2021:e13793.

Maess B, Schröger E, Widmann A. High-pass filters and baseline correction in M/EEG analysis-continued discussion. J Neurosci Methods 2016;266:171–2.

Mair P, Wilcox R. Robust statistical methods in R using the WRS2 package. Behav Res Methods 2020;52(2):464–88.

Mammone N, La Foresta F, Morabito FC. Automatic artifact rejection from multichannel scalp EEG by wavelet ICA. IEEE Sens. J. 2011;12(3):533–42.

Meltzer JA, Negishi M, Mayes LC, Constable RT. Individual differences in EEG theta and alpha dynamics during working memory correlate with fMRI responses across subjects. Clin Neurophysiol 2007;118(11):2419–36.

Miyakoshi M. Makoto's preprocessing pipeline. https://sccnucsdedu/wiki/Makoto%27s_preprocessing_pipeline Accessed February 2018;1:2019.

Mumtaz W, Rasheed S, Irfan A. Review of challenges associated with the EEG artifact removal methods. Biomed Signal Process Control 2021;68 102741.

Mutanen TP, Metsomaa J, Liljander S, Ilmoniemi RJ. Automatic and robust noise suppression in EEG and MEG: The SOUND algorithm. Neuroimage 2018;166:135–51.

Muthukumaraswamy S. High-frequency brain activity and muscle artifacts in MEG/EEG: a review and recommendations. Front Hum Neurosci 2013;7:138.

Nixon AE, Mazzola JJ, Bauer J, Krueger JR, Spector PE. Can work make you sick? A meta-analysis of the relationships between job stressors and physical symptoms. Work Stress 2011;25(1):1–22.

Nolan H, Whelan R, Reilly RB. FASTER: fully automated statistical thresholding for EEG artifact rejection. J Neurosci Methods 2010;192(1):152–62.

Oostenveld R, Fries P, Maris E, Schoffelen J-M. FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput Intell Neurosci 2011;2011 156869.

Palmer JA, Kreutz-Delgado K, Makeig S. AMICA: An adaptive mixture of independent component analyzers with shared components. Swartz Center for Computatonal Neursoscience. University of California San Diego, Tech Rep; .

Pavlov YG, Adamian N, Appelhoff S, Arvaneh M, Benwell CS, Beste C, et al. # EEGManyLabs: Investigating the replicability of influential EEG experiments. Cortex 2021;144:213–29.

Pion-Tonachini L, Kreutz-Delgado K, Makeig S. ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. Neuroimage 2019;198:181–97.

Raimondo F, Kamienkowski JE, Sigman M, Fernandez SD. CUDAICA: GPU optimization of infomax-ICA EEG analysis. Comput Intell Neurosci 2012;2012 206972.

Ranjan R, Sahana BC, Bhandari AK. Ocular artifact elimination from electroencephalography signals: a systematic review. Biocybern Biomed Eng 2021;41(3):960–96.

Robbins KA, Touryan J, Mullen T, Kothe C, Bigdely-Shamlo N. How sensitive are EEG results to preprocessing methods: a benchmarking study. IEEE Trans Neural Syst Rehabil Eng 2020;28(5):1081–90.

Rogasch NC, Sullivan C, Thomson RH, Rose NS, Bailey NW, Fitzgerald PB, et al. Analysing concurrent transcranial magnetic stimulation and electroencephalographic data: a review and introduction to the open-source TESA software. Neuroimage 2017;147:934–51.

Rogasch NC, Zipser C, Darmani G, Mutanen TP, Biabani M, Zrenner C, et al. The effects of NMDA receptor blockade on TMS-evoked EEG potentials from prefrontal and parietal cortex. Sci Rep 2020;10(1):1–12.

Roštáková Z, Rosipal R. Determining the number of components in the PARAFAC model with a nonnegative tensor structure: A simulated EEG data study. Neural Comput Appl 2022;34:14793–805.

Rousselet GA. Does filtering preclude us from studying ERP time-courses? Front Psychol 2012;3:131.

Smith JL, Jamadar S, Provost AL, Michie PT. Motor and non-motor inhibition in the Go/NoGo task: an ERP and fMRI study. Int J Psychophysiol 2013;87(3):244–53.

Somers B, Bertrand A. Removal of eye blink artifacts in wireless EEG sensor networks using reduced-bandwidth canonical correlation analysis. J. of Neural Eng. 2016;13(6) 066008.

Somers B, Francart T, Bertrand A. A generic EEG artifact removal algorithm based on the multi-channel Wiener filter. J. of Neural Eng. 2018;15(3) 036007.

Somers B, Francart T, Bertrand A. MWF toolbox for EEG artifact removal. https://githubcom/exporl/mwf-artifact-removal/blob/master/gui/GUI_manualpdf 2019.

Tanner D, Norton JJ, Morgan-Short K, Luck SJ. On high-pass filter artifacts (they're real) and baseline correction (it'sa good idea) in ERP/ERMF analysis. J Neurosci Methods 2016;266:166–70.

Tesche C, Uusitalo M, Ilmoniemi R, Huotilainen M, Kajola M, Salonen O. Signal-space projections of MEG data characterize both distributed and well-localized neuronal sources. Electroencephalogr Clin Neurophysiol 1995;95(3):189–200.

Urigüen JA, Garcia-Zapirain B. EEG artifact removal—state-of-the-art and guidelines. J Neural Eng 2015;12(3) 031001.

Uusitalo MA, Ilmoniemi RJ. Signal-space projection method for separating MEG or EEG into components. Med Biol Eng Comput 1997;35(2):135–40.

van Dijk H, van Wingen G, Denys D, Olbrich S, van Ruth R, Arns M. The two decades - Brainclinics research archive for insights in neurophysiology (TD-BRAIN) database. Sci Data 2022;9(1):1–10.

Winkler I, Debener S, Müller K-R, Tangermann M. On the influence of high-pass filtering on ICA-based artifact reduction in EEG-ERP. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. (EMBC): IEEE; 2015. p. 4101–5.

Zhang H, Zhao M, Wei C, Mantini D, Li Z, Liu Q. Eegdenoisenet: A benchmark dataset for deep learning solutions of eeg denoising. arXiv preprint arXiv:200911662 2020.

Zhang Q, van Vugt M, Borst JP, Anderson JR. Mapping working memory retrieval in space and in time: A combined electroencephalography and electrocorticography approach. Neuroimage 2018;174:472–84.