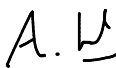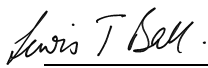# SKAO SCIENCE DATA PRODUCTS: A SUMMARY

SKA-TEL-SKO-0001818 Revision 02

Classification: UNRESTRICTED

Document type: NOT

Date: 2025-06-16

Status: RELEASED

Role	Name	Designation	Affiliation	Signature	Date
Author	Vinod Arumugam	Operations Scientist	SKAO	A. U	2025-06-16
Owner	Shari Breen	Head of Science Operations	SKAO	Shari Breen	2025-06-17
Approver	Antonio Chrysostomou	Deputy Director of Operations	SKAO	Antonio Chrysostomou	2025-06-16
Released by	Lewis Ball	Director of Operations	SKAO	Lewis T Ball	2025-06-16

FULL LIST OF AUTHORS

V. Arumugam, S. Breen, R. Bolton, A. Chrysostomou, D. Fenech, A. Noutsos, B. A. Pampliega, C. Smith

TABLE OF CONTENTS

LIST OF FIGURES

1 Introduction

1.1 Purpose of the document

This document provides a summary of the data products that SKA users can expect, as well as the processes through which they will be delivered. The information presented here is an updated version of that presented in the "Observatory Establishment and Delivery Plan" [RD1], with some additional details or emphasis appropriate for the provision of this brief reference document for the scientific community.

The information provided here will inform the community as users adapt to the SKAO workflow. The SKAO will provide users with science data products, processed according to the parameters and pipelines selected by the users prior to observations being executed. This is, however, not to say that users will not be able to interact with, and potentially amend the chosen Science Data Processor (SDP) pipelines or parameters as the observations for a large project begin. The SKAO recognises that there will sometimes be a need to observe a small fraction (i.e., a few hours to a few percent of the project depending on the individual requirements) of a large project and deliver the derived data products together with the calibrated visibility data to the SRCNet for consideration by the project PIs. Based on these data, the PIs may wish to test and fine tune the selected SDP pipelines (which will be accessible within the SRCNet) and ultimately amend the requested SDP workflow for the full project. Following this process, it is expected that the remainder of the project will be processed by the SDP without further interaction with the project PIs. This is a mutually beneficial process to ensure that the completed project satisfies the science goals in a timely and efficient manner.

While a similar process might be available for standard PI projects until an observing mode or capability has been fully verified, PIs of these projects should not expect a similar SDP feedback stage to be available to them following these initial stages of Operations.

1.2 Scope of the document

This document includes pertinent information about the provision and delivery of SKA data products to science users when the Observatory is in steady-state Operations, as well as the advanced analysis that users might conduct within the SKA Regional Centre (SRC) Network. As a subset of the information provided in [RD1], we refer the reader to that document for further information. The information provided here and in [RD1] provides the current plan for SKAO science data products and takes precedence over details provided in any document that may have previously been referenced within the community.

2 Data products

In general, the SKAO defines three types of SKA science data products split between two categories:

Observatory Data Products (ODPs):

Observation-level data products (OLDPs) are calibrated data products generated by SKAO pipelines and are based on data obtained from a single execution of a scheduling block (SB).

Project-level data products (PLDPs) are calibrated data products generated by combining several, related, observation-level data products, delivering the project requirements as outlined in the proposal.

Software pipelines to generate both OLDPs and PLDPs will be specified during the proposal stage. The Observatory is responsible for the generation of both types of ODPs, providing the workflows, software, conducting quality assessment (QA), reproducibility and that both product types are appropriately stored and

made available to users (i.e. archived). OLDPs will be generated by the SDP, but the generation of PLDPs will often require the combination of data taken at different epochs (e.g. for a deep integration or large mosaic) and will likely employ the use of SRCNet resources. Generation of ODPs of either type (OLDP or PLDP) will not require any interaction by the science users. In the case of large projects, it is acknowledged that it may be necessary to allow PIs to check the output of the selected SDP pipeline on some test data (with a feedback loop to adjust SDP pipeline parameters before going "full steam ahead"), or to provide multiple data products from the SDP (e.g. a set of different imaging parameters) whilst the visibility data are still available to establish the best parameters for future processing.

Advanced Data Products (ADPs):	ADPs are the user-generated products, produced through the detailed and rigorous analysis and modelling of Observatory data products (either at the observation or project level). The generation of ADPs will usually require some level of interactive visualisation and examination of data, as well as comparison to data from other SKA observations or other facilities.
	Science users are responsible for the generation of ADPs.

2.1   Pipelines and Data Products

OLDPs will be generated by SKAO workflows and pipelines specified at the proposal stage. There is no user interaction with the SDP (and other SKAO pipelines), but users will be able to define workflow parameters in their project SBs which contain definitions of "processing blocks", including technical details such as required spatial and spectral resolutions as well as continuum image bandwidth intervals. These will be set in advance of the observations being scheduled using the Observation Design Tool. A processing block is an atomic unit of data processing for the purposes of the SDP's internal scheduler. Each processing block will reference a processing workflow and each SB will indicate one or more processing blocks to be used (specifying, e.g., ingest, self-calibration, Data Product preparation).

The complete list of OLDPs that SDP will be capable of generating is detailed below. Each of these will have associated data processing logs and a QA report.

Image Products 1: Cubes	● Imaging data for continuum as cleaned and restored images. This includes a bandwidth-averaged multi-frequency synthesis (MFS) image, and a cube of evenly spaced channels covering the requested bandwidth to allow for in-band spectral-index estimation. The number of channels in the cube will be driven by the scientific requirements of the project, but is anticipated to be typically 10 channels. A spectral index map could also be generated if requested.
	● Residual image (i.e., residuals after applying CLEAN) in continuum.
	● Clean component image (or a table, which could be smaller).
	● Spectral line cube after (optional) continuum subtraction, where individual channels in the cube are imaged and deconvolved independently.
	● Residual spectral line image (i.e., residuals after applying CLEAN).
	● Representative point spread function for observations (cut-out, small in size compared to the field of view (FOV)).
	● Observations of standard pointed mosaics will include sensitivity (noise) cubes as a function of sky direction, polarisation, and frequency, to be used when constructing the larger mosaic.
	For continuum data products, the imaging products are expected to be generated with a wideband multi-frequency deconvolution algorithm similar to the

implementation in WSClean, called joined-channel deconvolution, as described in [RD2]. Briefly, in the joined-channel deconvolution method, the deconvolution is done over multiple frequency channels simultaneously, rather than on each channel independently. Given the broadband frequency range over which channels will be integrated, the intrinsic spectral variation of the sources and the varying primary beam response across the multiple channels needs to be accounted for – this is done by fitting a polynomial function determined by measuring the flux at the position of the clean components in each of the individual channels. In the case of the spectral line cubes, individual channels are imaged and deconvolved independently over the requested bandwidth of each spectral window. Unless specified otherwise, continuum and spectral line imaging products will include Stokes I, Q, U and V cubes.

Images can be generated at a higher cadence than the length of a single observing block, e.g. generating an image every 20 minutes from an observing block of 4 hours to probe the variability of a source. Note that this is not the same as the fast-imaging workflow for detecting transients, as described below in "Imaging Transient Source Catalogue".

Image Products 2: uv Grids	• Calibrated visibilities gridded at the spatial and frequency resolution required by the experiment. One grid per channel to be imaged and per facet[1] (the FFT of the dirty map of each facet) will be generated. • Accumulated weights for each uv cell in each grid (without additional weighting applied). The weights of each output channel are gridded on an individual grid, and in the case of multi-frequency channels, the weights of all input channels making up each frequency-integrated output channel will be gridded on one grid.
Wide area scanning and drift scanning data products	The wide area scanning (WAS) and drift scan observing modes will produce linear mosaics in each of the scheduled observations, as continuum or spectral line products. The larger mosaics combining the data from the different observing epochs will need to be generated as a PLDP. As will be the case for standard pointed mosaics, sensitivity (noise) cubes will be generated as a function of sky direction, polarisation, and frequency to be used when constructing the larger mosaic. The approach to be taken on how the data is combined (in the uv- or image-plane) is currently under investigation. Both cross- (interferometric) and auto-correlation (single-dish) data are likely to be recorded during the use of these modes.
Calibrated Visibilities	Calibrated visibility data and direction-dependent calibration information, with time and frequency averaging appropriate to the science case.[2]
Auto-correlation data	In specific cases, auto-correlation data i.e. data from single-dish observations can be delivered if requested as a data product. This ODP is under development, and discussions are being held with the community on the best approach of how they will be delivered.
Imaging Transient Source Catalogue	Time ordered catalogue of candidate transient objects produced by the SDP. This catalogue is generated from detection alerts from the real-time Fast Imaging

---

[1] Due to the wide field of view of the SKA telescopes, the field is divided into narrow facets to account for direction-dependent effects.

[2] a null calibration table with zero averaging could be applied to allow access to raw visibilities in exceptional circumstances.

pipeline, which is a dedicated process designed to search for slow radio transients with variability ranging from seconds to hours using snapshot images.

In the Fast Imaging pipeline, dirty Stokes I images, generated from a local sky model (LSM) subtracted visibility dataset, are searched for transient sources. No deconvolution is done as transient sources are expected to be unresolved and sparse. The extracted fluxes at the position of the detected sources are then recorded in the catalogue. Other quantities in the catalogue may include error estimates, a variability flag, parameters of the source fitting, etc. While the images generated by the Fast Imaging pipeline are not generally specified for saving – typically the maps are made, searched, and discarded, with the detections recorded in this catalogue – it might be possible to deliver a subset of these images, e.g. snapshots displaying source variability.

Pulsar Timing Solutions	For each detected pulsar, the output data from the pulsar timing section will include: the original input data (folded pulsar profiles that the SDP receives, before the PST pipelines on the SDP perform further RFI excision, calibration and averaging) as well as averaged versions of these data products (either averaged in polarisation, frequency, or time) in PSRFITS format; time of arrival (ToAs); residuals from the current best-fit timing model for the pulsar and updated ephemerides (timing models).
Detected filterbank archives (formerly dynamic spectrum)	High time- and frequency-resolution, full Stokes polarisation filterbank data, with configurable frequency and time resolutions, requantised to efficient bit depth (1, 2, 4, 8, 16 or 32 bits/sample). These data are not restricted to pulsar targets, with applications including e.g. exoplanet hunting, Solar studies, SETI.
Flow-through archives	Archive of compressed portions of the input tied-array beam signal from the beam former. Raw, dual-polarisation beamformed voltages requantised to appropriate bit depth (1, 2, 4, 8, 16 or 32 bits/sample), with some simple RFI mitigation; and a portion of the frequency band in one or both polarisations selected for archiving in PSRDADA format.
Sieved Pulsar and Transient Candidates	In pulsar acceleration search mode: an Optimised Candidate Lists & Data (OCLD), i.e. a list of parameters containing period, acceleration, signal-to-noise and dispersion measure of each of the detected candidates; and a single cube per candidate, representing the total intensity (Stokes I) of the signal for the folded and de-dispersed pulsar profiles (without information on polarization) will be provided.
	In single pulse search: a Single Pulse Optimised Candidate Lists & Data (SPOCLD) i.e. a list of parameters containing time of occurrence, signal-to-noise, dispersion measure and pulse width; and full polarisation (Stokes IQUV) filterbank data (at a configurable frequency and time resolution) covering the pulse duration. Additional metadata derived from candidate sifting across all beams and per-beam analysis using Machine Learning algorithms will be also included. Discovery of sufficiently interesting transients will generate an alert.
Transient Buffer Data	Raw, dual-polarisation voltage data captured when the transient buffer is triggered. The maximum bandwidth that can be buffered is limited to 150 MHz for SKA-Low and 400 MHz for SKA-Mid, and the bit depth of the data is 2 bits or better sampling – the bandwidth, bit depth and buffer dump time are configurable as a compromise

of each other, e.g. using a narrower bandwidth will allow for a higher bit depth or longer dump time to be used. Further details on the specifications of the transient buffer capture mode can be found in [AD1].

Science Alerts Catalogue	Catalogue of International Virtual Observatory Alliance (IVOA) formatted science alerts, produced and communicated by the SDP. This catalogue provides a searchable and retrievable record of past alerts.
Science Product Catalogue	Central, searchable repository/database for all processed ODPs, containing crucial scientific metadata that supports IVOA models. Produced by the SDP, it facilitates data discovery and access for users through interfaces like the SRCs, ensuring that successful queries directly lead to data delivery.
VLBI	VLBI products are the VLBI beams (beamformers at the CSP can combine the outputs from the antennas/stations to produce a stream of data — array beam — that will be used for VLBI observations), in VDIF format, to be correlated at an external VLBI correlator.
Ancillary data products	Ancillary data products are expected to include: ● Local Sky Model catalogue: a subset of the Global Sky Model (GSM) containing the sources relevant to other requested ODPs. These are the sources in the FOV, as well as, potentially, strong sources outside of the FOV of those ODPs. Initially, the LSM is filled from the GSM. During data processing, the sources found in the images are added to the LSM. The resultant LSM might be superior to the GSM (e.g. improved sensitivity or accounting for source variability), and can be used to provide an updated GSM (and subsequently an LSM) for future observations of the same field. The LSM catalogue may include details such as the sky coordinates, flux density at different frequencies, spectral index, etc. ● Science Product Catalogue: a central, searchable repository/database for all processed ODPs, containing crucial scientific metadata that supports IVOA models. Produced by the SDP, it facilitates data discovery and access for users through interfaces like the SRCs, ensuring that successful queries directly lead to data delivery. ● subset of the Telescope Model (e.g. antenna locations, names, geodetic coordinates, etc.) ● subset of the Telescope State information ● gain tables (calibration solutions) ● QA report ● processing block configuration (a representation of the processing done for the observations, e.g. workflow scripts to execute, compute resources required, etc.) ● processing logs ● Other required ObsCore Data Model elements (metadata required to interface with the IVOA)

Multiple data products can be produced from the same observation, limited only by scientific justification and resource availability. No combination of data products should be considered innately mutually exclusive. However, consideration and justification of the overall SDP processing load required for the generation of all products associated with a particular SB will be needed at the time of proposal assessment (as part of technical feasibility), but also at the project planning stage and as the SB is scheduled.

Delivery of raw visibility data as a data product (with or without averaging and/or calibration) is technically possible and is likely to be necessary in exceptional circumstances (perhaps less than 1% of projects in an observing cycle). However, in steady-state Operations, the SKAO is responsible for the delivery of calibrated data products and proposals requesting raw visibility data are expected to be very much the exception, and will require a detailed plan for calibration and the generation of data products.

2.1.1   Project-level Data products

For many projects it will be necessary to combine multiple related OLDPs to fulfil the science goals outlined in the observing proposal. Given the limited capacity of the SDP and the time between the execution of the respective SBs for a project, it is likely that many PLDPs will be generated within the SRCNet but will remain the responsibility of the SKAO. Like OLDPs, these will be created using SKAO workflows and will have associated data processing and QA information included in a log file. As the combination of OLDPs, PLDPs will necessarily be drawn from the same list of possible SDP data products given above.

2.2   Quality Assessment

All ODPs will have an associated QA report from the entire SKA processing chain. These QA reports will chiefly contain information generated by the SDP (on astrometry, photometry, radiometry, polarimetry and spectrometry), but will also contain other relevant information from the CSP (specifically in the PSS and PST) and the SKA-Low Monitoring, Control and Calibration System.

2.3   Lifecycle of Data Products

OLDPs generated within each SDP will be delivered to the SRCNet where users will access them. Once all planned SDP pipelines needing a particular set of raw data have been completed, those data will be deleted to free space in the SDP's buffer (along with any intermediate data products). In addition to being delivered to the SRCNet, all OLDPs will be stored in a long-term preservation (LTP) system for both telescopes. Once delivered to the SRCNet, users will be able to access these data products if they have appropriate permissions. As products move out of their proprietary periods, user access will be public. Some products may be public from the outset.

Execution of an SKA project may take anything from minutes to months, depending on the length of each scheduling block instance, the number of different scheduling blocks required and any special weather or observational conditions they may need. For each scheduling block instance actually executed, the SDP will ingest data and work on initial pipelines in real-time, but thereafter data is placed into a "cold" buffer for storage until the bulk SDP processing generating the OLDPs can occur - this will depend on the loading of the SDP. For example, the most challenging processing, resulting from a long (e.g., 10 hour) SB, may take several times longer than the SB duration to be processed by the SDP, thus data processing will sit in a queue until such a time that the SDP batch processing system has the capacity to process a pending job. Based on current estimates, this means that it could take a couple of weeks after SB execution for OLDPs to be generated by the SDP. Once OLDPs are created by the SDP they will be queued for delivery to the SRCNet. Again, this process is a bottleneck that depends on the available bandwidth, so further delays in delivery are possible and should be expected if high data volume projects are undertaken. (The metric to bear in mind here is that the anticipated connectivity of

each SDP into the SRCNet is 100 Gbit/s, or ~1 petabyte per day. Capacity on this link is a resource that will need to be planned for alongside storage and compute capabilities.) The delivery of OLDPs to the SRCNet, is an area that will need to be prototyped extensively, and timescales analysed, during the construction and science verification periods.

As each OLDP is generated and made available in the SRCNet, PIs and Co-Is will have access to them in that form, but full PLDPs (if applicable to a project), that need to be generated by the SKAO using SRCNet resources, can only be created after all necessary OLDPs are in place. As the generation of the PLDPs will be the last stage before the PIs can access their final requested data products, it will be important that there is minimal delay in their creation (subject to compute availability), and that user expectations on these timescales are well managed.

2.4 Roles and Responsibilities of SKAO, SRCNet and the SKA User in the generation of data products

In Figure 1, a 'swimlane' diagram is used to show where the responsibilities of the SKAO, the SRCNet, and the SKA user (whether as a PI/Co-I, or an archive user) lie with respect to delivering the SKA science programme. It shows two phases: the project execution phase and the science extraction phase. The SKAO is responsible for the project execution phase, which includes the generation, calibration, and delivery of ODPs into the SRCNet. The SRCNet is then responsible for supporting the SKA community of users in extracting the science from the ODPs delivered to them for publication and dissemination.

The boundary between each 'lane' reflects the need to strike a balance between centralisation and the desire for software quality and data traceability, and the ability to declare that the Observatory's responsibilities to a specific PI or Key Science Project (KSP) team have been achieved, against the somewhat competing need to encourage scientific freedom and innovation.

The science extraction phase will generally result in the generation of further, more advanced data products (ADPs) as a consequence of the advanced analysis and modelling that will be employed by the science community. Those ADPs that will appear in publications, or that will be made public, will be added to the SKA science archive and made available to all users (while respecting the appropriate proprietary access periods). This will raise efficiency in the SRCNet by avoiding repetition of the processing to generate those products.

It should be noted that while this figure highlights the ultimate responsibility for each of the data products, it doesn't fully capture all of the important roles played in their generation. Specifically, the generation of PLDPs are listed as the responsibility of the SKAO since we will provide and maintain the software pipelines to generate them from OLDPs, but the actual processing to generate them will primarily be undertaken within the SRCNet.

The generation of the OLDPs and ADPs are somewhat more simply presented in Figure 1; the generation of OLDPs are the sole responsibility of the SKAO and the generation of ADPs are the joint responsibility of the users and SRCNet.

2.5 Data Management Model for the SKAO and the SRCNet

Once generated, ODPs will be delivered to the SRCNet, with copies retained in an LTP system at SDP sites for backup purposes. The LTP will be a high-latency data storage system, ensuring data preservation in case SRCNet copies are lost. The LTP is a last-resort backup, and will not be an actively managed storage element within the SRCNet.

Within SRCNet, a global data management service will oversee storage and access. Instead of each SRC node maintaining full copies of all data, a global data management service will govern data storage, access, and transfers between SRC nodes. This service helps balance storage demands by tagging excess copies for potential deletion, ensuring nodes can clear space when necessary.
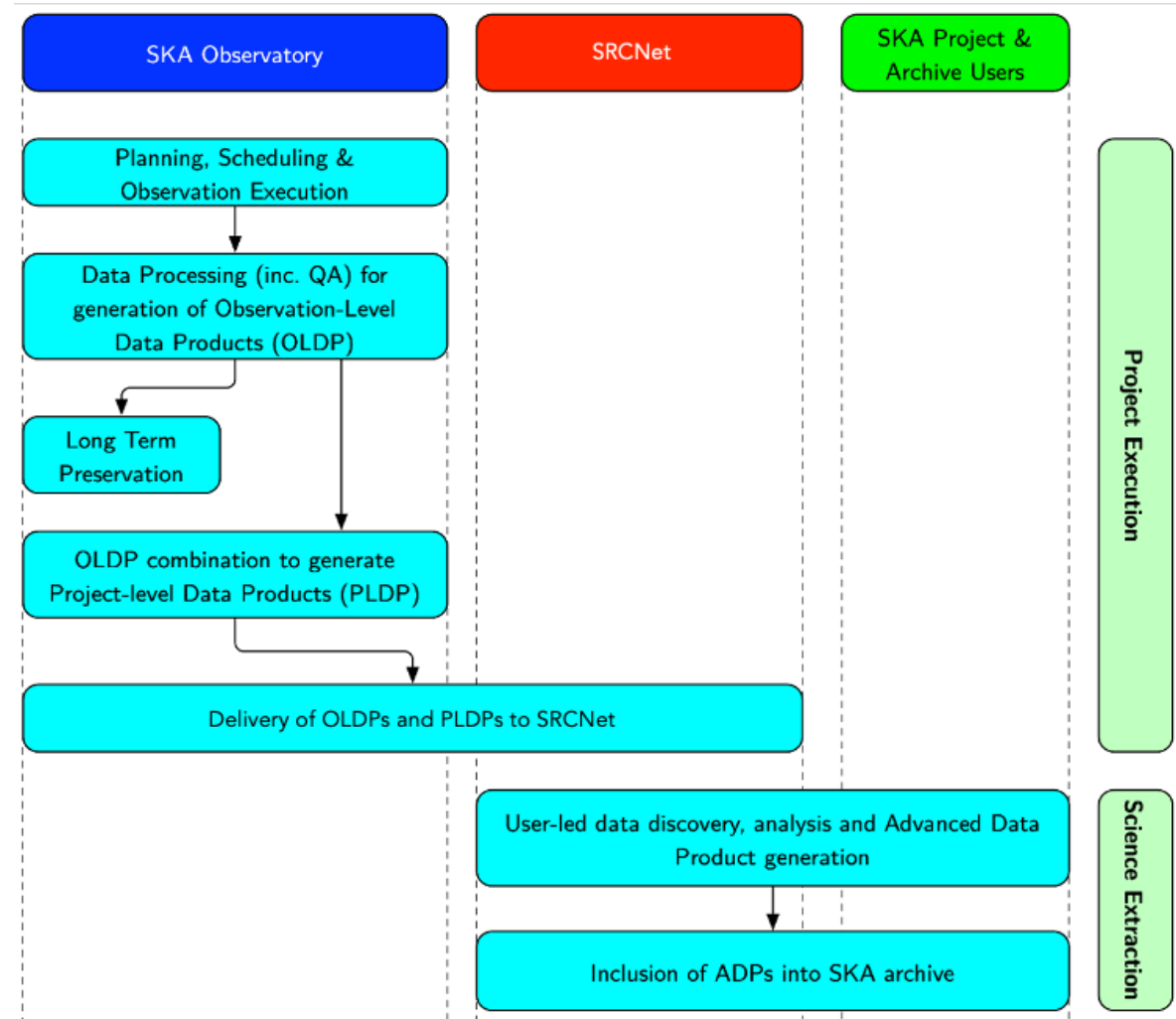

Figure 1. Swimlane diagram showing the responsibilities for the generation of science data products during the project execution and science extraction phases of a science project. The left lane depicts the Observatory's responsibilities, the middle lane those of the SRCNet, and the right lane for the PIs and Co-Is of SKA projects and general archive users. Observation-Level and project-level data products are generated by the Observatory, while advanced data products are produced by users within the SRCNet.

A References

A.1 Applicable Documents

The following documents are applicable to the extent stated herein. In the event of conflict between the contents of the applicable documents and this document, **the applicable documents** shall take precedence.

[AD1] SKAO-TEL-0002665, A year in the life of the SKA telescopes: overview and main outcomes, Rev01

A.2 Reference Documents

The following documents are referenced in this document. In the event of conflict between the contents of the referenced documents and this document, **this document** shall take precedence.

[RD1] SKA-TEL-SKO-0001722, SKAO Establishment & Delivery Plan, Rev03

[RD2] An optimized algorithm for multiscale wideband deconvolution of radio astronomical images, Offringa and Smirnov, 2017

LIST OF ABBREVIATIONS

AD	Applicable Document
ADP	Advanced Data Products
CSP	Central Signal Processor
FFT	Fast Fourier Transform
GSM	Global Sky Model
IVOA	International Virtual Observatory Alliance
KSP	Key Science Project
LSM	Local Sky Model
LTP	Long Term Preservation
OCLD	Optimised Candidate Lists & Data
ODP	Observatory Data Product
OLDP	Observation-level Data product
PI	Principal Investigator
PLDP	Project-level Data Product
PSS	Pulsar Search
PST	Pulsar Timing
QA	Quality Assessment
RD	Reference Document
RFI	Radio Frequency Interference
SB	Scheduling Block
SDP	Science Data Processor
SKA	Square Kilometre Array
SKAO	SKA Observatory
SPOCLD	Single Pulse Optimised Candidate Lists & Data
SRC	SKA Regional Centre
SWG	Science Working Group
ToA	Time of Arrival
VDIF	VLBI Data Interchange Format
VLBI	Very Long Baseline Interferometry
WAS	Wide Area Scanning

DOCUMENT HISTORY

Revision	Date Of Issue	Engineering Change Number	Comments
A	2021-04-07		First draft release for internal review
B	2021-04-28		Revisions including feedback from Operations staff and members of the SKA Regional Centre Steering Committee WGs
C	2021-05-11		Final comments from all signatories
01	2021-05-15		1st Release
02	2025-06-16		2nd Release

DOCUMENT SOFTWARE

	Package	Version	Filename
Word processor	MS Word	Office 365	SKA-TEL-SKO-0001818-02_ScienceDataProducts_A Summary
Block diagrams			
Other			

ORGANISATION DETAILS

Name	SKA Observatory
Registered Address	Jodrell Bank Lower Withington Macclesfield Cheshire, SK11 9FT, UK
Fax	+44 (0)147 777 3400
Website	www.skao.int